École des Hautes Études en Sciences Sociales

Doctoral school nº 465 Économie Panthéon Sorbonne

Paris-Jourdan Sciences Économiques

PH.D THESIS

Field: Economics

SIMON LÖWE

Essays in empirical economics

Thesis supervised by: Luc Behaghel and Thomas Breda

Date of defense: November 7th, 2023

Referees:	Pauline GIVORD, Head of the department of economic studies at INSEE
	Richard MURPHY, Assistant Professor at University of Texas Austin
Jury:	Simon JÄGER, Associate Professor at Massachusetts Institute of Technology (MIT)
	Marc GURGAND, Chaired Professor at Paris School of Economics (PSE)
Supervisor:	Luc BEHAGHEL, Associate Professor at Paris School of Economics
Cosupervisor:	Thomas BREDA, Full-time junior researcher at CNRS

École des Hautes Études en Sciences Sociales

École doctorale nº 465 Économie Panthéon Sorbonne

Paris-Jourdan Sciences Économiques

DOCTORAT

Discipline : Économie

SIMON LÖWE

Essais en économie empirique

Thèse dirigée par : Luc Behaghel et Thomas Breda

Date de soutenance : le 7 novembre 2023

Rapporteurs:	Pauline GIVORD, Cheffe du département des études économiques à l'INSEE
	Richard MURPHY, Assistant Professor à University of Texas Austin
Jury:	Simon JÄGER, Associate Professor au Massachusetts Institute of Technology (MIT)
	Marc GURGAND, Chaired Professor à l'École Économique de Paris
Superviseur:	Luc BEHAGHEL, Maître de Conférences à l'École Économique de Paris
Cosuperviseur:	Thomas BREDA, Chargé de recherche au CNRS

Acknowledgements

I want to thank my advisors Luc Behaghel and Thomas Breda for supporting me in this endeavor and for convincing me to start it in the first place. I also want to thank Xavier D'Haultfoeuille for giving his time and help when Yagan Hazard and I decided to try ourselves at econometrics. I also want to thank Simon Jäger for giving me the opportunity to do a stimulating exchange to MIT and for agreeing to be on the jury. I also want to thank Thomas Le Barbanchon for trusting me to work with him during my fourth year, and allowing me to discover and live in yet another country. Finally, thank you to Pauline Givord, Richard Murphy and Marc Gurgand for agreeing to read, referee and judge this thesis.

I'm also grateful for the companions along the way that made doing and finishing this thesis possible. Special mention goes to Yagan Hazard who has been my coauthor and friend since the very beginning of my somewhat adventurous conversion from physicist to economist. Thanks to the friends and office mates, Leonard Bocquet, Sophie Cottet, Louise de Gaudemaris, Nikki Kerkogozou, Shakked Noy, Eric Teschke, for both discussing the PhD with me and distracting me from PhD.

I obviously would not be where I am without my parents who have always supported my admittedly very long life in academia so far. Thank you for your unconditional support. Thank you for helping me figure out that economics was something I wanted to do.

Finally, Emily, my love, thank you for absolutely everything. I can say with certainty that none of this would have been possible without you. You are always there for me. You read everything I wrote, discussed everything I wanted to discuss, pushed me when I needed it, helped me to allow myself to rest and convinced me that I'm maybe not as stupid as I generally think I am. Without these I wouldn't have been able to do this. But even more importantly, thank

you for giving me what ultimately really matters, which is most definitely not this collection of words. This PhD wouldn't have been possible (I know I've said this a lot but it is true) without the knowledge that even if I failed at it, I would still would be infinitely happy, because I get to live and share my life with you.

Contents

Short Summary

This thesis examines 3 different subjects in empirical economics. The first chapter proposes a methodology to improve estimation in experiments with imperfect compliance. The second chapter documents the effect of stigma on take-up of benefits. Finally, the third chapter studies the effect of fixed-term contracts on wages, focusing on differential rent-sharing and compensating wage differentials.

Chapter 1

The evaluation of many policies of interest (e.g., educational and training programs) inevitably face incomplete treatment group take-up. Estimation of causal effects in these controlled or natural "experiments with imperfect compliance" usually relies on an Instrumental Variable (IV) strategy, which often yields imprecise and thus possibly uninformative inference when compliance rates are low. We tackle this problem by proposing a Test-and-Select estimator that exploits covariate information to restrict estimation to a subpopulation with non-zero compliance. We derive the asymptotic properties of our proposed estimator under standard and weak-IV-like asymptotics, and study its finite sample properties in Monte Carlo simulations. We provide conditions under which it dominates the usual 2SLS estimator in terms of precision. Under an assumption on the degree of treatment effect heterogeneity, our estimator remains first-order unbiased with respect to the Local Average Treatment Effect (LATE) estimand, setting it apart from alternatives in the burgeoning literature on the use of first-stage heterogeneity to improve the precision of IV estimators. This robustness to treatment effect heterogeneity and the potential for precision gains are illustrated using Monte Carlo simulations and two

empirical applications. Applying this new estimation procedure to the returns to schooling example (where compulsory schooling laws serve as instruments for educational attainment), we document that our methodology reduces standard errors by 12% to 48% depending on specifications.

Chapter 2

Media may change economic behaviour by stigmatizing certain actions. Focusing on English tabloid newspapers, we study the impact of stories with negative sentiments towards benefit claimants on the take-up of social benefits. Using an event study approach, we find a 4-5% reduction in applications for a social benefit in the three days following the publication of such a story. For individuals who would have otherwise applied for benefits this equates to an estimated loss of at least £36 (€42, $45) per person in response to a single newspaper story. This effect is more pronounced in areas with lower incomes and higher rates of benefit application. We argue this reduction is driven by stories increasing the salience of welfare stigma. These results suggest that media can have a direct effect on economic behaviour and that stigma is a significant driver of non-take-up of benefits.

Chapter 3

The effect of fixed-term contracts on wages is theoretically ambiguous. I introduce a simple monopsonistic model which incorporates segmented fixed-term and open-ended contracts markets, differential rent-sharing and compensating wage differentials for job security. Using exhaustive French administrative data, I then confirm several predictions of the model, documenting novel empirical facts about FTCs and wages. While the overall average wage gap between the contract types is estimated to be a precise zero, wages vary differentially between both contract types, resulting in a wage premium for low productivity firms and a wage penalty for high productivity firms. I then further illustrate the differential rent-sharing mechanism by documenting that firms only extend about half of the wage premium of open-ended contract

workers to fixed-term contract workers. Finally, I show that the degree of rent-sharing varies with local labor market concentration for OECs but not for FTCs.

Résumé court

Cette thèse examine trois sujets différents dans le domaine de l'économie empirique. Le premier chapitre propose une méthodologie pour améliorer l'estimation dans les expériences avec "imperfect compliance". Le deuxième chapitre documente l'effet de la stigmatisation sur le recours aux prestations sociales. Enfin, le troisième chapitre étudie l'effet des contrats à durée déterminée sur les salaires, en se concentrant sur le partage différentiel de la rente et la compensation salariale associée à l'insécurité de l'emploi.

Chapitre 1

L'évaluation de nombreuses politiques d'intérêt (par exemple, les programmes d'éducation et de formation) est inévitablement confrontée à une participation incomplète du groupe de traitement. L'estimation des effets causaux dans ces "expériences avec imperfect compliance" contrôlées ou naturelles repose généralement sur une stratégie de variable instrumentale (IV), qui produit souvent une inférence imprécise et donc potentiellement non informative lorsque les taux de conformité sont faibles. Je m'attaque à ce problème en proposant un estimateur Test-and-Select qui exploite l'information de variables observables pour restreindre l'estimation à une sous-population dont le taux de "compliance" n'est pas nul. Je déduis les propriétés asymptotiques de l'estimateur proposé dans le cadre d'une asymptotique standard et d'une asymptotique de type IV faible, et j'étudie ses propriétés en échantillon fini dans le cadre de simulations de Monte Carlo. Je fournis des conditions dans lesquelles notre estimateur domine l'estimateur 2SLS habituel en termes de précision. Sous une hypothèse sur le degré d'hétérogénéité de l'effet de traitement, notre estimateur reste sans biais de premier ordre par rapport à l'estimand standard

de Local Average Treatment Effect (LATE), ce qui le distingue des alternatives dans la littérature sur l'utilisation de l'hétérogénéité de first stage pour améliorer la précision des estimateurs IV. Cette robustesse à l'hétérogénéité de l'effet de traitement et le potentiel de gain de précision sont illustrés par des simulations de Monte Carlo et deux applications empiriques. En appliquant cette nouvelle procédure d'estimation à l'exemple des rendements de la scolarité (où les lois sur la scolarité obligatoire servent d'instruments pour le niveau d'éducation), je montre que ma méthodologie réduit les erreurs standard de 12% à 48% selon les spécifications.

Chapitre 2

Les médias peuvent modifier les comportements économiques en stigmatisant certaines actions. En nous concentrant sur les tabloïds anglais, nous étudions l'impact d'articles exprimant des sentiments négatifs à l'égard des demandeurs de prestations sociales. En utilisant une approche "event study", nous constatons une réduction de 4 à 5% des demandes de prestations sociales dans les trois jours suivant la publication d'un tel article. Pour les personnes qui auraient autrement demandé des prestations, cela équivaut à une perte estimée à au moins £36 (€42, $45) par personne en réponse à un seul article de journal. Cet effet est plus prononcé dans les régions où les revenus sont plus faibles et les taux de recours plus élevés. Nous affirmons que cette réduction est due au fait que les articles augmentent la saillance de la stigmatisation liée au recours aux prestations sociales. Ces résultats suggèrent que les médias peuvent avoir un effet direct sur le comportement économique et que la stigmatisation est un facteur important de non-recours aux prestations.

Chapitre 3

L'effet des contrats à durée déterminée sur les salaires est théoriquement ambigu. Je présente un modèle de monopsone simple qui intègre des marchés segmentés de contrats à durée déterminée et indéterminée, un partage différentiel de la rente et une compensation salariale associée à l'insécurité de l'emploi. À l'aide de données administratives françaises exhaustives, je confirme

ensuite plusieurs prédictions du modèle, en mettant en évidence des faits empiriques inédits concernant les CDDs et les salaires. Alors que l'écart salarial moyen global entre les types de contrats est estimé à zéro, les salaires varient de manière différentielle entre les deux types de contrats, ce qui se traduit par une prime salariale pour les entreprises à faible productivité et une pénalité salariale pour les entreprises à forte productivité. J'illustre ensuite le mécanisme de partage différentiel de la rente en montrant que les entreprises n'étendent aux travailleurs sous contrat à durée déterminée qu'environ la moitié de l'avantage salarial dont bénéficient les travailleurs sous contrat à durée indéterminée. Enfin, je montre que le degré de partage de la rente varie en fonction de la concentration du marché du travail local pour les CDI, mais pas pour les CDD.

Résumé long

Chapitre 1

Les stratégies de variables instrumentales (IV) font partie intégrante de la boîte à outils standard des économistes appliqués et des chercheurs en sciences sociales. Cela est dû en partie à leur utilisation pour l'estimation des effets causaux dans des expériences contrôlées ou naturelles avec "imperfect compliance". Ces expériences sont omniprésentes dans la recherche appliquée, car de nombreuses interventions (telles que les programmes d'éducation ou de formation) ne peuvent être imposées à un groupe sélectionné au hasard. Dans ce cas, les membres du groupe de traitement sont simplement encouragés ou ont la possibilité de bénéficier de l'intervention. Pourtant, l'estimation IV dans ces contextes est souvent affectée par de faibles taux de compliance, ce qui conduit à une variance accrue et donc à une inférence potentiellement non informative sur les effets causaux d'intérêt. Compte tenu de l'investissement financier et humain considérable associé à la mise en œuvre d'un essai contrôlé randomisé (ECR) typique et de la rareté des expériences naturelles existantes, le fait de ne pas informer les décideurs politiques en raison de l'imprécision des procédures d'estimation dans ces expériences a un coût social important.

Cependant, un faible taux de conformité *moyen* peut masquer des comportements de compliance très hétérogènes dans des sous-populations présentant des caractéristiques observables différentes. Les chercheurs ont donc la possibilité d'améliorer la précision de leurs estimations en tenant compte de cette hétérogénéité. Dans cet article, nous proposons et étudions les propriétés d'une méthode intuitive permettant de tirer parti de cette hétérogénéité. Notre estimateur Test-and-Select restreint l'estimation IV aux sous-populations ayant des taux de compliance significatifs non nuls dans l'échantillon. En excluant de l'échantillon d'estimation les sous-groupes

dont on estime que l'effet de première étape est nul, on se débarrasse des observations qui n'apportent que peu ou pas de signal sur l'effet causal en question, mais ajoutant éventuellement un bruit considérable à la distribution de l'estimateur IV standard.[1].

Le papier est structuré comme suit. Nous soulignons tout d'abord les pièges de la mise en œuvre "naïve" d'une règle de sélection basée sur les taux de compliance estimés, puis nous proposons qu'une procédure de "data-splitting" (fractionnement des données) apporte une solution simple à ce problème. Ensuite, nous étudions les propriétés asymptotiques de l'estimateur Test-and-Select dans le cadre de séquences asymptotiques standard et de type "weak-IV". La première analyse nous permet d'illustrer les gains potentiels en termes de précision, tandis que la seconde vise à mieux approcher les propriétés d'échantillon fini de l'estimateur que nous proposons. Ces analyses soulignent la robustesse de la procédure Test-and-Select à l'hétérogénéité de l'effet de traitement. En effet, nous montrons qu'elle reste sans biais au premier ordre pour l'effet causal habituel qui nous intéresse — communément appelé "Local Average Treatment Effect" (LATE) — dans des situations d'hétérogénéité de l'effet de traitement qui généreraient un biais au premier ordre dans les stratégies d'estimation alternatives proposées dans la littérature. Enfin, nous étudions les propriétés de cet estimateur sur échantillon fini dans des simulations de Monte-Carlo et dans deux applications — une expérience naturelle utilisant des changements dans les lois sur la scolarité obligatoire comme instrument pour l'éducation, et une expérience à grande échelle sur le conseil en matière de recherche d'emploi. Ces sections illustrent (i) les gains potentiels de précision résultant de la mise en œuvre de notre méthodologie au lieu de l'estimateur 2SLS habituel, et (ii) la meilleure robustesse de notre estimateur à l'hétérogénéité de l'effet de traitement par rapport aux autres solutions.

Le fardeau que représente un faible taux de compliance pour la précision de l'estimateur Two-Stage-Least-Squares (2SLS) est bien connu pour la plupart des empiristes, et la formule de variance de l'estimateur 2SLS dans le cas simple où la variance des erreurs (notée σ_ε^2) est

[1]Nous utiliserons de manière équivalente les termes "taux de compliance" et "first stage" dans le présent document. En effet, dans le modèle IV "simple" avec un instrument binaire et un traitement binaire considéré ici, le coefficient de première étape - c'est-à-dire le coefficient de l'instrument issu de la régression de l'indicateur de traitement sur l'indicateur de l'instrument — coïncide avec la part d'observateurs dans la (sous-)population sur laquelle le modèle IV est estimé

homoscédastique en est la meilleure illustration.[2]En désignant par N la taille de l'échantillon, p la part des individus encouragés et π la part des compliers, nous obtenons:

$$\text{Var}\left[\widehat{LATE}^{2SLS}\right] = \frac{1}{N} \cdot \frac{1}{\pi^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)}$$

Ici, nous pouvons clairement remarquer qu'un faible taux de compliance a un effet disproportionné sur la variance de l'estimateur 2SLS du LATE. Prenons un exemple illustratif, en étudiant la variance de deux expériences évaluant le même programme, l'une avec un taux de compliance de 10% ($\pi = 0,1$) et l'autre avec un taux de compliance parfaite ($\pi = 1$). Le taux de compliance dans la première expérience n'est que 10 fois plus faible que dans la seconde, et pourtant la taille de l'échantillon doit être 100 fois plus importante pour atteindre la même précision (variance) que dans l'autre expérience. En d'autres termes, supposons qu'il soit possible, dans la première expérience, d'utiliser certaines variables observables pour identifier la sous-population des compliers. En se concentrant sur cette fraction (10%) de la population, on diviserait l'échantillon d'estimation par 10, mais on diminuerait la variance d'un facteur 10, ce qui améliorerait considérablement l'inférence. En résumé, même si une expérience donnée passe avec succès certains tests d'identification faibles — ce qu'elle pourrait faire même avec des taux de compliance relativement faibles — un faible taux de compliance peut encore être très nuisible en réduisant considérablement la précision, conduisant éventuellement à une inférence non informative.

Pour fixer les idées dans un cadre plus concret, considérons l'instrument du trimestre de naissance (Angrist and Krueger, 1991). Cet instrument repose sur l'idée qu'en raison des lois sur la scolarité obligatoire, les enfants nés en début d'année seront légalement autorisés à abandonner leurs études plus tôt que ceux nés en fin d'année — ce qui conduit les premiers à effectuer moins d'années d'études que les seconds en moyenne. Cependant, les préférences en matière d'éducation sont susceptibles d'être très hétérogènes en fonction de multiples dimensions (par exemple, le revenu et les qualifications des parents). Par exemple, il se peut qu'aucun des

[2]Ici, ε est le terme d'erreur structurel dans ce que l'on appelle généralement l'équation de la "second stage", c'est-à-dire la régression du résultat sur la variable de traitement (et sur certains contrôles si nécessaire).

enfants de parents appartenant aux 50% (ou 60, 70, 80%) supérieurs de la distribution des revenus n'envisage jamais d'abandonner l'école avant d'en avoir légalement le droit. Dans ce cas, leur trimestre de naissance n'aurait aucun effet sur leur niveau d'éducation. En bref, certaines sous-populations pourraient ne pas réagir à l'instrument du trimestre de naissance et, de ce fait, ne contribueraient pas à l'identification du LATE. Il est important de noter que l'existence de ces groupes de non-compliers ne constitue pas une menace pour l'identification,[3] mais leur présence dans l'échantillon d'estimation réduit la précision avec laquelle le LATE est estimé. Il est intuitif d'exclure ces groupes de l'échantillon d'estimation. Ce chapitre montre comment rendre cette stratégie opérationnelle et étudie ses propriétés.

Sous des hypothèses d'"asymptotique standard", qui conduit en fin de compte à une sélection parfaite des groupes sans compliers, notre estimateur cible le même paramètre LATE que l'estimateur 2SLS/Wald habituel, tout en apportant des gains de précision. Cependant, de telles asymptotiques sont susceptibles de fournir une mauvaise approximation du comportement de l'estimateur que nous proposons dans des échantillons finis. C'est pourquoi nous étudions des séquences asymptotiques plus réalistes où les taux de compliers sont autorisés à être "locaux à zéro" dans certains groupes, parce que de telles asymptotiques laissent de la place pour des exclusions erronées de groupes avec une part non nulle de personnes ayant respecté le traitement. En l'absence d'hypothèses sur l'hétérogénéité de l'effet de traitement, l'estimateur que nous proposons présente un biais de premier ordre pour le LATE, car les groupes exclus à tort peuvent avoir un effet de traitement arbitrairement important.

Nous fournissons donc des conditions sous lesquelles l'estimand que notre méthodologie cible est équivalent au premier ordre à l'estimand LATE. Une condition suffisante pour que cette propriété soit remplie est de limiter le degré d'hétérogénéité de l'effet de traitement entre les groupes au même ordre de grandeur que la variation d'échantillonnage. En d'autres termes, l'hétérogénéité entre les groupes est telle qu'elle ne serait pas systématiquement détectée dans des échantillons finis. Nous expliquons en détail pourquoi il s'agit d'une condition raisonnable en pratique. Nous proposons également une stratégie de data-splitting (et cross-

[3]Pour être précis, ces groupes de non-compliers ne menacent pas l'identification à moins qu'ils ne représentent la majorité de l'échantillon. Dans ce cas, le LATE peut être *faiblement* identifié.

fitting) qui génère une inférence valide malgré le pré-test sur lequel repose notre stratégie d'estimation. Nous étudions les propriétés d'échantillon fini de la procédure proposée dans des simulations de Monte Carlo. Un paquetage R `late.rest` qui met en œuvre notre estimateur (et permet la réplication de nos simulations Monte-Carlo) est disponible à `https://github.com/simon-lowe/late.rest`.

Chapitre 2

Les médias ont souvent été accusés de stigmatiser certains groupes de la société. Cela peut potentiellement affecter leur comportement économique dans le monde réel, si prendre une certaine décision risque d'accroître cette stigmatisation. Dans cet article, nous nous concentrons sur un groupe stigmatisé - les bénéficiaires de prestations sociales - et nous nous demandons si les médias stigmatisants influencent la décision des personnes éligibles de bénéficier des prestations sociales.

Le non-recours est un problème omniprésent, le taux de recours à certaines prestations sociales étant estimé à moins de 30% dans certains pays européens (Dubois et al., eds, 2015). Des travaux antérieurs montrent que les personnes les plus pauvres et les plus marginalisées sont moins susceptibles de faire recours aux prestations (Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2019), ce qui fait du non-recours un obstacle sérieux au ciblage de l'aide sociale. Cependant, alors qu'une multitude de données qualitatives et d'enquêtes attestent de l'existence d'une stigmatisation autour des prestations sociales (Baumberg, 2016; Morrison, 2019), il existe très peu de données causales sur la question de savoir si la stigmatisation affecte réellement le recours aux prestations sociales.

Compte tenu de ce manque de connaissances, ce chapitre étudie l'effet de la stigmatisation des bénéficiaires de prestations par les médias sur le recours aux prestations sociales.

Nous nous concentrons sur les articles de presse à contenu négatif sur les bénéficiaires de prestations du journal "tabloïd" britannique, *The Sun*. Au cours de la période étudiée, il s'agissait du journal ayant le plus fort tirage au Royaume-Uni, ainsi que de celui dont la

couverture médiatique était la plus négative à l'égard des bénéficiaires de prestations.[4] Pour collecter ces articles, nous avons entraîné un réseau de neurones profond à classifier si les articles traitent de prestations sociales ou non. Nous l'utilisons pour extraire tous les articles du *Sun* sur la période 2013-2019 qui sont sur ce sujet.[5] Parmi ces articles, nous avons ensuite identifié manuellement ceux qui contenaient des sentiments négatifs explicites à l'égard des personnes qui reçoivent ou demandent des prestations. Les articles ont tendance à se regrouper dans le temps, souvent avec un article d'actualité un jour suivi d'un commentaire le même jour ou les jours suivants. Nous appelons ces regroupements des "suites d'articles".

Nous fournissons des preuves causales que ces suites d'articles ont conduit à une diminution de l'utilisation des prestations sociales. Nous faisons correspondre des données quotidiennes sur le nombre de demandes de Universal Credit, une prestation composite soumise à conditions de ressources pour les personnes en âge de travailler, avec les dates du premier article d'une suite d'articles contenant une couverture négative des bénéficiaires de prestations. Notre stratégie empirique exploite l'exogénéité locale plausible du moment de la publication des articles de presse. Il est peu probable que la publication d'un article sur les bénéficiaires de prestations soit indépendante des tendances à long terme en matière de recours. Cependant, nous soutenons qu'elle est exogène au nombre de personnes demandant des allocations un jour donné, par rapport au jour précédent. En utilisant une approche "event study", nous examinons si les demandes de Universal Credit sont affectées par la publication d'un article contenant une couverture négative des bénéficiaires de prestations.

Nous constatons une baisse importante et significative du nombre de demandes de Universal Credit dans les jours qui suivent la publication d'une telle suite d'article. En moyenne, le nombre de demandes diminue de 4 à 5% dans les trois jours suivant la publication. Ces résultats sont tirés par les régions géographiques où les revenus sont les plus faibles et où les taux de demande

[4]Turn2us, qui fait partie de l'organisation caritative de lutte contre la pauvreté Elizabeth Finn, a documenté en 2012 que *The Sun* a la couverture médiatique la plus négative à l'égard des bénéficiaires de prestations de tous les journaux britanniques. `https://www.theguardian.com/news/datablog/2012/nov/20/benefits-stigma-newspapers-report-welfare`.

[5]Pour 2013-2016, les résumés des articles sont disponibles dans ProQuest, ce qui nous permet d'exécuter ce classificateur. Après 2017, ProQuest a cessé de traiter les résumés et seuls les titres sont disponibles. Par conséquent, nous utilisons des recherches par mots clés sur LexisNexis (qui ne nous permet pas de mettre en œuvre notre classificateur) pour extraire les articles de cette période.

d'allocations sont les plus élevés. Nous trouvons également des preuves suggestives que les résultats sont tirés par les articles placés plus près du début du journal (déterminé par le numéro de page). Nous ne constatons aucune différence significative dans le nombre de demandes à la suite d'un article neutre ou positif à l'égard des bénéficiaires de prestations. Ces résultats sont robustes à une variété de spécifications et à l'exclusion d'un événement particulier.

Nous construisons un cadre conceptuel de l'utilisation de l'aide sociale afin d'explorer les mécanismes qui sous-tendent ces résultats. Ce cadre modélise la décision de recours à l'aide sociale en l'associant à divers "mauvais" types. Les bénéficiaires de prestations sociales ont été associés à de nombreux types "mauvais", notamment les personnes peu compétentes, les resquilleurs (Friedrichsen et al., 2018), les pauvres (Holford, 2015), et les fraudeurs (Gavin, 2021). De nombreux articles de presse de notre échantillon font référence à ces types. Les suites d'articles peuvent affecter la probabilité d'être considéré comme l'un de ces types, la pénalité pour être considéré comme l'un de ces types, ou la perception de ces paramètres par l'individu. Les coûts de l'image sociale sont modélisés à l'aide de ces paramètres avec un groupe de référence social, et l'image de soi est modélisée de manière analogue avec le soi comme groupe de référence, conformément à la littérature sur le "self-signalling" (Bodner and Prelec, 2002; Mijovic-Prelec and Prelec, 2010; Bénabou et al., 2018). La décision de participation dépend alors de ces coûts liés à l'image sociale et à l'image de soi, ainsi que d'autres coûts et avantages associés à la demande de prestations.

Nous soutenons que le lien entre la couverture négative dans les journaux et le non-recours est dû aux coûts liés à l'image de soi et à l'image sociale, que nous appelons conjointement le "coût de la stigmatisation". La couverture négative renforce l'association entre les "mauvais" types et la décision de s'inscrire, ainsi que la pénalité liée au fait d'être associé à l'un de ces types. Nous avançons l'hypothèse qu'une grande partie de l'effet est due à la saillance accrue de ces paramètres sous-jacents à court terme, plutôt qu'aux fluctuations des paramètres eux-mêmes.

Nous excluons d'autres mécanismes susceptibles d'influer sur les autres coûts ou avantages (non liés à la stigmatisation) dans la décision de recours. Le non-recours aux prestations sociales est généralement attribué au manque d'informations sur les conditions d'éligibilité (Bhargava

and Manoli, 2015; Anders and Rafkin, 2022) et aux coûts liés à la complexité de la demande (Finkelstein and Notowidigdo, 2019). L'effet que nous observons ne peut pas être dû à l'apport d'informations sur les prestations, car cela conduirait à une augmentation des demandes plutôt qu'à la diminution que nous observons. Les suites d'articles pourraient également fournir des informations sur le processus de demande et modifier la perception qu'ont les demandeurs potentiels des coûts de complexité de la demande. Cela pousserait l'effet dans la direction que nous observons, mais nous constatons qu'aucun des articles de notre échantillon ne discute du processus de demande, de l'administration du Universal Credit ou de problèmes liés aux demandes, ce qui rend ce mécanisme peu probable. En revanche, de nombreux articles traitent de cas de fraude aux prestations. Nous démontrons que la dissuasion de la fraude aux prestations ne peut expliquer l'ampleur de nos résultats.

Nous constatons que le taux de demande revient aux niveaux antérieurs au traitement après cinq jours, sans augmentation détectable des demandes pour compenser celles qui n'ont pas été faites directement après une suite d'articles négative. Il semble donc qu'une suite d'articles peut être associée à quatre jours de traitement, en moyenne, les personnes n'étant plus traitées le cinquième jour. Cette interprétation confirme le mécanisme de saillance décrit ci-dessus. Il est plausible qu'après une période de quatre jours, l'importance accrue de la pénalité se dissipe, car d'autres nouvelles ou d'autres sujets deviennent plus importants. Cela n'exclut pas un décalage du recours à plus moyen terme de quelques semaines. Toutefois, ce phénomène n'est pas inhabituel pour les facteurs déterminant de non-recours. Une grande partie du non-recours consiste à retarder la demande plutôt qu'à ne jamais la faire.

Étant donné que les demandes de prestations sociales sont associées à un revenu monétaire direct, cela nous permet de mettre un coût direct sur la volonté de payer pour éviter la stigmatisation, pour les "compliers". En 2021, le montant mensuel minimum du crédit universel était de £368 (€429, $462) pour une personne seule, soit environ £12 (€14, $15) par jour.[6]Si nous supposons, de manière conservative, que les personnes retardent simplement leur demande de Univresal Credit de trois jours, un calcul sommaire donne une perte de £36 (€42, $45) par personne, en réponse à une seule suite d'article. Il s'agit d'une limite inférieure car nous ne

[6]Voir `https://www.gov.uk/universal-credit/what-youll-get`

constatons pas d'augmentation compensatoire des demandes après trois jours, ce qui implique que les personnes retardent leur demande plus longtemps ou ne la déposent jamais. Si une suite d'articles dissuade complètement une personne de déposer une demande, le coût s'élèverait alors à £810 (€944, $1,017), soit le montant moyen du crédit universel versé en 2021.[7]

Ce chapitre contribue à plusieurs littératures. Tout d'abord, notre recherche jette un nouvel éclairage sur les facteurs de non-recours aux prestations sociales. De nombreuses données qualitatives attestent de l'existence de la stigmatisation liée à l'aide sociale. Baumberg (2016) montre que 20,4% des Britanniques pensent que les gens devraient avoir honte de demander au moins une prestation, et 16,9% ont déclaré qu'ils auraient honte de faire recours aux prestations sociales.[8]. Cependant, les résultats des enquêtes sur la stigmatisation sont difficiles à interpréter d'un point de vue causal, car l'admission de la stigmatisation est elle-même stigmatisante, et lorsque de multiples facteurs affectent une décision, il est difficile pour les individus d'indiquer l'importance de l'un d'entre eux en particulier. Celhay, Meyer and Mittag (2021) ont constaté que les résultats d'enquête sur la perception de prestations sont particulièrement problématiques. Friedrichsen, König and Schmacker (2018) abordent ces questions en étudiant l'effet de la stigmatisation en laboratoire, et constatent un effet significatif sur le recours hypothétique à des prestations. Nous fournissons, à notre connaissance, la première preuve causale réelle de l'effet de la stigmatisation sur le recours aux prestations.

En outre, nous nous appuyons sur la littérature plus large sur le non-recours, qui a révélé que le manque d'informations (Holford, 2015; Bhargava and Manoli, 2015; Anders and Rafkin, 2022) et la complexité des procédures de demande (Finkelstein and Notowidigdo, 2019) sont des facteurs importants. Une autre partie de la littérature examine le rôle de la procrastination dans le non-recours, en particulier dans le contexte des subventions à l'éducation (Dynarski, 2007; Sunstein, 2013; Narayan, 2020).Toutefois, si ces articles reconnaissent l'importance de la stigmatisation, ils ont du mal à l'isoler des autres facteurs de non-recours. Par exemple, Holford

[7]Voir `https://www.gov.uk/government/statistics/universal-credit-statistics-29-april-2013-to-universal-credit-statistics-29-april-2013-to-14-july-2022`

[8]Des enquêtes similaires ont été menées dans d'autres contextes. Aux États-Unis, Stuber and Kronebusch (2004) trouve que plus de la moitié des personnes interrogées sont d'accord avec l'affirmation suivante : "De nombreuses personnes bénéficiant d'une aide sociale ne veulent pas que les autres sachent qu'elles en bénéficient" et entre 35 et 45% pensent que "dans ce pays, les personnes bénéficiant d'une aide sociale sont paresseuses"

(2015) trouvent des effets de pairs sur l'utilisation des repas scolaires gratuits, mais concluent qu'il s'agit principalement d'un effet d'information.[9] Dans notre cadre empirique, tout effet de l'information agirait dans la direction opposée à l'effet de la stigmatisation, ce qui nous permet d'isoler cet effet de manière convaincante.

Deuxièmement, notre travail contribue à la littérature en plein essor sur les effets de l'image sociale et de l'image de soi sur le comportement économique (voir Bursztyn and Jensen (2017) pour une revue), ainsi qu'à l'importante littérature en sociologie et en psychologie sociale qui étudie et conceptualise la stigmatisation (eg. Link and Phelan, 2001; Major and O'Brien, 2005; Pescosolido and Martin, 2015). Une série d'essais contrôlés randomisés étudient l'effet de la stigmatisation sur des résultats aussi divers que la participation électorale aux États-Unis (Dellavigna et al., 2017), le comportement d'épargne des travailleurs du sexe en Inde (Ghosal et al., 2022), et l'effort au lycée (Bursztyn, Egorov and Jensen, 2019a). L'article le plus similaire, Osman and Speer (2023), examine la participation à une formation professionnelle et à un salon de l'emploi en Égypte, après des interventions visant à atténuer les préoccupations liées à la stigmatisation. Ils constatent que leurs interventions diminuent en fait la participation, ce qu'ils interprètent comme la preuve d'un effet de stigmatisation. À notre connaissance, nous sommes les premiers à montrer un effet de stigmatisation dans une expérience naturelle. Nous contribuons en outre à la littérature sur l'image de soi et l'image sociale en disposant d'un moyen naturel d'"évaluer" le coût de ces préoccupations, étant donné que le non-recours se traduit directement par le renoncement à un montant monétaire spécifique.

Enfin, nous contribuons à la littérature sur la persuasion par les médias. Un certain nombre d'articles examinent la manière dont les médias peuvent façonner les attitudes à l'égard des groupes marginalisés. Djourelova (2023) a récemment étudié l'effet sur les attitudes envers les immigrants aux États-Unis après que l'Associated Press a interdit l'utilisation du terme "immigrant illégal", constatant des changements significatifs dans le soutien aux politiques d'immigration. Ivandic, Kirchmaier and Machin (2019) montrent comment la couverture médiatique augmente le nombre de crimes de haine islamophobes au Royaume-Uni après des attentats

[9] De même, Bhargava and Manoli (2015) mènent une petite expérience parallèle sur la stigmatisation et trouvent des effets négatifs sur l'utilisation, mais ils concluent que cela est probablement dû à un changement dans les perceptions de la complexité.

terroristes.[10]DellaVigna, Enikolopov, Mironova, Petrova and Zhuravskaya (2014) montrent que la radio favorise les attitudes nationalistes à la frontière serbo-croate. Alors que ces articles se concentrent sur la réaction du groupe majoritaire, nous nous concentrons sur la réaction des individus marginalisés ou potentiellement marginalisés. Ce faisant, nous complétons une vaste littérature sociologique qui a examiné le rôle des médias dans la stigmatisation de la pauvreté et de l'aide sociale (voir par exemple Morrison, 2019).

De manière plus générale, la littérature sur la persuasion par les médias s'est concentrée sur les effets de l'exposition à un média particulier (DellaVigna and Kaplan, 2007; Enikolopov et al., 2011; Levy, 2021). Notamment, Foos and Bischof (2021) se concentre également sur l'exposition à *The Sun*, en examinant les effets sur l'euroscepticisme. Dans cet article, nous ouvrons cette boîte noire, en nous concentrant sur les effets au niveau microéconomique de l'exposition à des articles individuels.

Chapitre 3

L'un des thèmes centraux des débats sur la politique européenne du marché du travail au cours des trois dernières décennies a été la détermination du niveau approprié de la législation sur la protection de l'emploi, avec un accent particulier sur la réglementation des contrats à durée déterminée (CDD) et des contrats à durée indéterminée (CDI). Alors que les discussions académiques ont principalement tourné autour de l'emploi et des flux du marché du travail, ainsi que des effets des CDD sur la productivité, moins d'attention a été accordée à leur impact sur la structure des salaires.[11]

L'effet théorique des CDD sur les salaires est ambigu. Certaines théories suggèrent que les CDD pourraient faire baisser les salaires en raison de la réduction du pouvoir de négociation, tandis que d'autres affirment que les travailleurs devraient être indemnisés pour la diminution

[10]Bursztyn, Egorov, Enikolopov and Petrova (2019b); Müller and Schwarz (2021, 2023) constatent également que les médias sociaux peuvent jouer un rôle de médiateur entre les opinions xénophobes et les crimes de haine.

[11]Il est intéressant de noter que cette question a également été soulevée dans une récente affaire du tribunal du travail allemand, où il a été décidé que des tâches égales devaient être rémunérées de manière égale, quel que soit le type de contrat (voir par exemple `https://www.spiegel.de/karriere/bundesarbeitsgericht-minijobber-muessen-den-gleichen-stundenlohn-bekommen-a-88f26250-208a-44` en allemand).

de la sécurité de l'emploi. Il est essentiel de comprendre les mécanismes qui déterminent la formation des salaires en présence de CDD pour avoir une vue d'ensemble de leur rôle sur le marché du travail et des effets potentiels des réformes.

Pour éclairer ce débat, cet article présente un modèle de monopsone simple avec marchés de travail segmentés, catégorisés par type de contrat, complété par une compensation salariale associée à l'insécurité de l'emploi. Le modèle prédit plusieurs nouveaux faits que je documente ensuite en utilisant des données administratives complètes de la France, ce qui distingue cette étude d'une grande partie de la littérature existante qui s'appuie sur des données agrégées en "cross-section".

En introduisant un arbitrage entre une compensation salariale associée à l'insécurité de l'emploi et le partage différentiel de la rente entre les CDD et les CDI, mon modèle fait quatre prédictions. Les deux premières prédictions concernent l'écart salarial au sein de l'entreprise entre les CDI et les CDD. Premièrement, il prédit que les entreprises à faible productivité versent des salaires plus élevés aux CDD qu'aux CDI, alors que c'est l'inverse pour les entreprises à forte productivité. Cela s'explique par le fait que la compensation salariale pour l'insécurité de l'emploi est constante, mais que le partage de la rente augmente avec la productivité. Les entreprises à faible productivité paient alors de faibles salaires à la fois aux CDI et aux CDD, mais elles doivent compenser les CDD pour la moindre sécurité de l'emploi, ce qui entraîne un salaire plus élevé pour les CDD dans les entreprises à faible productivité. La deuxième prédiction est que le signe moyen de l'écart salarial entre CDI et CDD est ambigu. Il s'agit d'un corollaire direct de la première prédiction, puisqu'il dépend en particulier de la distribution de la productivité des entreprises dans l'économie.

Les deux prédictions suivantes concernent le partage différentiel des rentes. La théorie prédit que les entreprises partagent moins les rentes avec les travailleurs en CDD qu'avec les travailleurs en CDI. Ce point a déjà été abordé dans la littérature, mais généralement dans un modèle de négociation. Dans le cas présent, il s'agit d'un mécanisme purement monopsonistique. De plus, le modèle prédit aussi que le degré de partage des rentes varie en fonction des élasticités de l'offre de travail.

Je commence mon analyse empirique en examinant l'écart salarial moyen entre CDD et CDI pour les salaires d'embauche. Je me concentre sur les salaires d'embauche pour éviter les complexités supplémentaires qu'apportent les dynamiques d'évolution de salaires. Après avoir contrôlé les effets de l'âge, du sexe, de la profession, du secteur et de l'entreprise, je documente que l'écart des salaires d'embauche entre les CDD et les CDI est précisément estimé à zéro. Cela suggère que des théories prédisant des écarts salariaux à la fois positifs et négatifs sont nécessaires pour comprendre l'effet des CDD sur les salaires. Pour mieux différencier ces théories concurrentes, je présente plusieurs nouveaux résultats empiriques.

Étant donné que mon modèle met l'accent sur le rôle de l'entreprise, j'analyse ensuite l'écart salarial entre les CDD et les CDI au niveau de l'entreprise. Je montre que les salaires en CDI et les CDD varient en fonction de la valeur ajoutée par travailleur avec des pentes différentes comme prévu par le modèle, générant une prime salariale pour les CDD dans les entreprises à faible productivité et une pénalité salariale dans les entreprises à forte productivité. Pour documenter le mécanisme de compensation salariale pour l'insécurité de l'emploi, je montre que les salaires des CDD diminuent avec la durée moyenne des CDD dans les entreprises.

J'analyse ensuite plus en détail le mécanisme de partage différentiel de la rente. En utilisant l'approche introduit par Card et al. (2016) pour étudier l'écart salarial entre hommes et femmes, et récemment appliquée à la littérature sur l'externalisation par Drenik et al. (2022), je compare les primes salariales de l'entreprise pour les CDD et les CDI séparément, qui sont identifiées à partir des déplacements des travailleurs entre entreprises(Abowd et al., 1999). Cela permet de mesurer la mesure dans laquelle les entreprises à hauts salaires pour les CDD sont également des entreprises à hauts salaires pour les CDI.

Conformément à l'hypothèse d'un degré de partage de la rente moindre pour les travailleurs en CDD, je montre qu'en moyenne, les entreprises qui offrent une prime salariale de 10% aux travailleurs en CDI n'offrent qu'une prime de 5.1% aux travailleurs en CDD au cours de la période 2010-2014. Ce pourcentage est très proche de la prime de 4.9% observée par Drenik et al. (2022) dans le contexte de l'externalisation en Argentine. Ces résultats sont conformes aux prédictions du modèle de monopsone concernant le partage différentiel de la rente.

Enfin, j'explore la dernière prédiction sur la variation du partage de la rente en utilisant la variation de la concentration de l'embauche par type de contrat par marché de travail local. Premièrement, je montre que les marchés CDD sont plus concentrés que les marchés CDI. Ensuite, je montre que pour les CDI, la pente de la valeur ajoutée salariale (et donc le partage de la rente) diminue à mesure que la concentration des CDI augmente, comme le prédit le modèle. Enfin, je montre que la concentration en CDD semble affecter très peu la pente de la valeur ajoutée salariale pour les CDD, ce qui suggère que la concentration est une source de pouvoir de marché moins importante pour les CDD que pour les CDI.

Cet article contribue à la littérature sur les CDD en présentant un point de vue monopsonistique qui met l'accent sur le rôle de l'entreprise et le partage différentiel de la rente. Je montre que les CDD jouent un rôle important pour comprendre le pouvoir de marche dans les marches de travail dual. De plus, je documente qu'il est important de considérer les effets de compensation salariale pour l'insécurité de l'emploi dans l'étude des salaires en CDD. Enfin, je documente de nouveaux faits empiriques en accord avec les considérations précédentes. L'objectif est de motiver des recherches plus approfondies, à la fois théoriques et empiriques, sur les effets des CDD sur la structure des salaires.

Chapter 1

Improving LATE estimation in experiments with imperfect compliance

Joint work with ***Yagan Hazard***

1 Introduction

Instrumental variables (IV) strategies are an integral part of the standard toolkit of applied economists and social scientists. This is due in part to their use for the estimation of causal effects in controlled or natural experiments with imperfect compliance. Such experiments are pervasive in applied research, since many interventions (such as education or training programs) cannot be imposed on a randomly selected group. Instead, in such cases, members of the treatment group are simply encouraged or given the opportunity to benefit from the intervention. Yet IV estimation in these settings is commonly plagued by low compliance rates, which lead to an inflated variance and thus possibly uninformative inference on the causal effects of interest.[1] Given the substantial financial and human investment associated with implementing a typical Randomized Controlled Trial (RCT) and the scarcity of existing natural experiments, failing to inform policymaking due to imprecise estimation procedures in such experiments has a

[1] By "uninformative inference", we mean for instance confidence intervals wide enough to include values that researchers (and policy-makers) would deem large enough to justify the implementation of the treatment at hand, and at the same time, values too low to lead to such conclusion.

significant social cost.

Yet a low *average* compliance rate can obscure highly heterogeneous compliance behaviors across sub-populations with different observable characteristics. This leaves room for researchers to improve the precision of their estimation by taking into account this heterogeneity. In this paper, we propose and study the properties of an intuitive way to take advantage of such heterogeneity. Our Test-and-Select estimator restricts IV estimation to sub-populations with significant non-zero compliance rates in sample. Excluding sub-groups estimated to have a zero first-stage effect from the estimation sample gets rid of observations that bring little to no signal on the causal effect of interest while possibly adding considerable noise to the distribution of the standard IV estimator.[2]

The present paper is structured as follows. We first underline the pitfalls of "naively" implementing such a selection rule based on estimated compliance rates, and then propose that data-splitting provides a simple fix to this issue. Next, we study the asymptotic properties of the Test-and-Select estimator under both standard and weak-IV-like asymptotic sequences. The former analysis allows us to illustrate the potential gains in precision while the latter aims at better approximating the finite sample properties of our proposed estimator. These analyses underline the robustness of the Test-and-Select procedure to treatment effect heterogeneity. Indeed, we show that it remains first-order unbiased for the usual causal effect of interest — commonly known as the Local Average Treatment Effect (LATE) — under patterns of treatment effect heterogeneity that would generate a first-order bias in alternative estimation strategies proposed in the literature. Lastly, we study the finite sample properties of this estimator in Monte-Carlo simulations and in two applications — a natural experiment using changes in compulsory schooling laws as an instrument for education, and a large-scale experiment on job search counseling. These sections illustrate (i) the potential gains in precision from implementing our methodology instead of the usual 2SLS estimator, and (ii) the improved robustness of our estimator to treatment effect heterogeneity compared to alternatives.

[2]We will use equivalently the terms "compliance rates" and "first-stages" in this paper. This is because in the "simple" IV model with a binary instrument and binary treatment considered here, the first-stage coefficient — i.e., the coefficient on the instrument from the regression of the treatment indicator on the instrument indicator — coincides with the share of compliers in the (sub-)population on which the IV model is estimated.

The burden placed by low compliance rate on the precision of the Two-Stage-Least-Squares (2SLS) estimator is well-known to most empiricists, and best illustrated by the variance formula of the 2SLS estimator in the simple case where the variance of the errors (denoted σ_ε^2) is homoscedastic.[3] Denoting by N the sample size, p the share of encouraged individuals, and π the share of compliers, we get:[4]

$$\text{Var}\left[\widehat{LATE}^{2SLS}\right] = \frac{1}{N} \cdot \frac{1}{\pi^2} \cdot \frac{\sigma_\varepsilon^2}{p \cdot (1-p)}$$

Here, we can clearly notice that a low compliance rate has a disproportionately large effect on the variance of the 2SLS estimator of the LATE. Let's take an illustrative example, studying the variance in two experiments evaluating the same program, one with a 10% compliance rate ($\pi = 0.1$) and another with a perfect compliance ($\pi = 1$). The compliance rate in the first experiment is only 10 times lower than in the second experiment, and yet the sample size needs to be a 100 times larger to reach the same precision (variance) as in the perfect compliance experiment. Put differently, suppose it were possible in the first experiment to use some observables to identify the subpopulation of compliers. Focusing on this fraction (10%) of the population would divide the estimation sample by 10, but it would still *decrease* the variance by a factor of 10, and thus significantly improve inference. In summary, even if a given experiment passes some weak identification tests successfully — which it could even with relatively low compliance rates — a low take-up rate can still be highly detrimental by immensely decreasing precision, possibly leading to uninformative inference.

[3] Here, ε is the structural error term in what is usually called the "second stage" equation, i.e., the regression of the outcome on the treatment variable (and some controls if necessary).

[4] For the unaccustomed reader, p and π might seem similar. Instead, p is the share of individuals who are incentivized (or assigned) to take the treatment, while π is the difference in *effective* treatment take-up rates between individuals who are encouraged to take the treatment and those who are not. These objects are defined more formally in section 2 after introducing our formal framework.
Meanwhile, the variance formula presented above can be derived from the standard 2SLS variance formula in the homoscedastic case. Indeed, denoting by D the (binary) endogenous variable, and Z the (binary) instrument, recall that the formula for the asymptotic variance of the 2SLS estimator is given by:

$$V^{2SLS} = \sigma_\varepsilon^2 \left[\Sigma_{DZ}\Sigma_{ZZ}^{-1}\Sigma_{ZD}\right]^{-1} = \frac{\sigma_\varepsilon^2}{\left(E[D|Z=1] - E[D|Z=0]\right)^2 \cdot E[Z](1-E[Z])} = \frac{\sigma_\varepsilon^2}{\pi^2 \cdot p(1-p)}$$

where we used the notation $\Sigma_{XY} \equiv E[XY']$ and the definitions $\pi = E[D|Z=1] - E[D|Z=0]$ and $p = E[Z]$.

To fix ideas in a more concrete setting, consider the quarter-of-birth instrument (Angrist and Krueger, 1991). This paper builds on the idea that because of compulsory schooling laws, children born at the beginning of the year will be legally allowed to drop out earlier than those born in the end of the year — which leads the former to complete fewer years of schooling than the latter on average. Yet preferences for education are likely to be highly heterogeneous along multiple dimensions (e.g., parent's income and qualifications). For instance, it could be that none of the children of parents belonging to the top 50% (or 60, 70, 80%) of the income distribution ever consider dropping out of school before being legally able to do so. In such a case, their quarter of birth would have no effect on their educational attainment. To put it briefly, some sub-populations might not react to the quarter-of-birth instrument, and as such they would not contribute to the identification of the LATE. Importantly, the existence of such non-compliant groups is not a threat to identification,[5] but their presence in the estimation sample does reduce the precision with which the LATE is estimated. It is intuitive to drop these groups without compliers from the estimation sample. This paper shows how to make this strategy operational and studies its properties.

Under "standard asymptotics",[6] which ultimately leads to a perfect selection of groups without compliers, our estimator targets the same LATE parameter as the usual 2SLS/Wald estimator, while yielding precision gains. Yet such asymptotics are likely to provide a poor approximation for the behavior of our proposed estimator in finite samples. Therefore we study more realistic asymptotic sequences where compliance rates are allowed to be "local-to-zero" in some groups,[7] because such asymptotics leave room for erroneous exclusions of groups with a non-zero share of compliers. Under no assumptions on treatment effect heterogeneity, our proposed estimator has a first-order bias for the LATE, as wrongly excluded groups could have

[5]To be precise, such non-compliant groups do not threaten identification unless they represent the majority of the sample. In such a case, the LATE might be *weakly* identified.

[6]The precise definition of what we call "standard asymptotics" is given in section 3.1.

[7]Compared to the weak instrument literature, in which such "local-to-zero" first-stages were first introduced (Staiger and Stock, 1997), we still maintain the assumption that the overall first-stage is well separated from 0, allowing strong identification of the LATE. In other words, we do not assume away the possibility that some sub-populations would have "local-to-zero"/weak first-stages, yet there must be at the same time some other groups in which first-stages are strong for our assumption to be satisfied.

an arbitrarily large treatment effect.[8] We thus provide conditions under which the estimand that our methodology targets is first-order equivalent to the LATE estimand. A sufficient condition for this property to be fulfilled is to restrict the degree of treatment effect heterogeneity across groups to be of the same order of magnitude as the sampling variation. In other words, between-group heterogeneity is such that it would not be systematically detected in finite samples. We discuss in detail why this is a reasonable condition in practice. We also propose a data-splitting (and cross-fitting) strategy that generates valid inference despite the pre-test our estimation strategy relies on. We investigate the finite sample properties of our proposed procedure in Monte Carlo simulations. An R-package `late.rest` that implements our estimator (and allows for replication of our Monte-Carlo simulations) is available at `https://github.com/simon-lowe/late.rest`.[9]

Related literature. For ethical reasons, many programs of interests (e.g., training programs) cannot be imposed on (or refused to) a random population of individuals. In the absence of any natural source of randomness in the allocation of the treatment of interest, evaluators can only use so-called encouragement designs in which a contrast in the program take-up rate between treated and control populations is created by randomly allocating incentives to take up the treatment, rather than randomly allocating the treatment itself. The seminal work of Angrist, Imbens and Rubin in a series of papers (Imbens and Angrist, 1994; Angrist et al., 1996) clarified the causal parameter — often called the Local Average Treatment Effect (LATE) — that can be identified from such controlled or natural experiments where an encouragement is used as an instrument for treatment. This parameter is "local" in the sense that it corresponds to the average treatment effect among the "compliers", that is the population for whom treatment status is

[8] An estimator $\hat{\theta}$ of a parameter θ has a "first-order" or "asymptotic" bias when the limiting distribution of $\sqrt{n} \cdot \left(\hat{\theta} - \theta\right)$ is not centered on 0. For instance, if $\hat{\theta}$ is asymptotically normal with first-order bias B, then we have: $\sqrt{n} \cdot \left(\hat{\theta} - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(B, \Sigma)$. Notice that it does not prevent $\hat{\theta}$ from being a consistent estimator of θ. Yet it indicates that it does not converge towards θ at a $\sqrt{n}$-rate, which can invalidate inference based on such asymptotic approximation. Notice that throughout the paper, we will use the term "$\sqrt{n}$-rate consistency" as synonymous to asymptotic unbiasedness. We acknowledge the fact that often times, these terms are not used equivalently, as a $\sqrt{n}$-rate consistent estimator can denote the situation where we have: $(\hat{\theta})\theta) = O(1/\sqrt{n})$. When used in this way, $\hat{\theta}$ can still be asymptotically biased while $\sqrt{n}$-rate consistent.

[9] As of 09/27/2023, we are still working on the optimization and development of the package, but it already contains a functioning implementation of the estimator presented in this paper.

affected by the randomly modified incentive.[10]

This better understanding of instrumental variables (IV henceforth) led to a large body of work on the limitations of such an identification strategy when the instrument is "weak", i.e., when it creates only little variation in the treatment of interest — in the language of Angrist et al. (1996), when where there are only very few compliers. Yet apart from weak identification issues, low compliance settings raise other challenges in particular by affecting the precision of IV estimators. And even though some important work has been done on the optimal choice among many (weak) instruments (Belloni et al., 2012; Hansen and Kozbur, 2012), it typically does not deal with heterogeneous treatment effects — a framework that has become predominant in analyses of RCTs with imperfect compliance since the aforementioned work of Angrist and Imbens. This might sound surprising as besides the weak identification issue, a low compliance rate also has a tremendous cost in terms of variance of the usual IV/2SLS estimator for the LATE.

Though still less developed than the weak instrument corpus, a burgeoning literature has revisited the IV estimation strategy to achieve precision gains when the first-stage is heterogeneous along observable covariates (Huntington-Klein, 2020; Coussens and Spiess, 2021; Abadie et al., 2022). This renewed interest can be related to empirical attempts to identify treatment effects "on those who take it up" (Crépon et al., 2015). Recently, Coussens and Spiess (2021) and Abadie et al. (2022) proposed to use a "weighted-IV" or "interacted-IV" estimator that is optimal in the constant treatment effect regime which, under treatment effect heterogeneity, identifies a convex weighted average of LATEs (Huntington-Klein, 2020). They illustrate the precision gains from using this estimator in the presence of first-stage heterogeneity along observables. Though the decrease in variance obtained using our estimator comes from a similar source, we differ by maintaining the goalpost of estimating the standard LATE parameter instead of a weighted average of LATEs.[11] We do so because the LATE parameter might be considered a more directly policy-relevant parameter, as it corresponds to an existing sub-population that can

[10] Notice that the same holds for natural experiments that would generate randomness in an *encouragement* to take up the treatment of interest. For instance, compulsory law schools create higher incentives for some individuals to attend school for a longer period of time, depending on their birthdate. In such settings, one can only recover the average "local" effect among compliers, who are the individuals who actually attended school for a longer period of time *because* of their birthdate and the associated compulsory schooling laws.

[11] In the words of Huntington-Klein (2020), the parameter targeted by such estimation strategy is a "Super-Local Average Treatment Effect", since it gives a disproportionate weight to groups with larger compliance rates.

be targeted by policy-makers by using the exact same encouragement device (instrument) as in the experimental setting. Ultimately, choosing between our approach and one in the vein of Coussens and Spiess (2021) or Abadie et al. (2022) boils down to choosing between (i) smaller variance gains yet limited deviations from the original estimand of interest, the LATE, or (ii) larger variance gains at the cost of potentially large changes in the targeted estimand. Our paper is also related to the literature on semiparametrically efficient estimation of the LATE parameter. A common assumption in this literature is that all groups (defined by observable covariates) in the population have a share of compliers well separated from 0.[12] Under that assumption, this literature characterizes the semiparametric efficiency bound and provides estimators reaching it. Yet such a bound does not apply when compliance rates can be 0 or local to 0 in some sub-groups defined by the covariates.[13] Our work also uses a two-step procedure akin to the one studied in Abadie et al. (2022), dropping groups of observations displaying non-significant first-stage coefficients prior to the estimation step. Yet Abadie et al. (2022) consider such a strategy mainly as a way to reduce a weak-IV issue, while the estimator (and its associated MSE minimization problem) they study afterwards heavily rely on their constant treatment effect assumption.[14]

The remainder of the paper unfolds as follows. Section 2 presents the general framework and introduces the proposed estimator. Section 3 develops the theoretical results, and section 4 suggests some extensions. Section 5 studies the finite sample properties of our proposed estimation strategy, and compares it to alternatives. Section 6 presents two empirical applications — the first on a natural experiment using variation in compulsory schooling laws as an instrument for educational attainment, and the second on a large-scale RCT on job search counseling. Lastly, section 7 concludes and presents some avenues for further research on this topic.

[12] See, e.g., assumption 1.ii in Hong and Nekipelov (2010) and assumption 1.iv in Singh and Sun (2021)

[13] Intuitively, the results from this literature do not apply in this case as identification of the LATE conditional on covariates — which is always assumed in this literature — fails for some values of the covariates.

[14] Abadie et al. (2022) do propose an interpretation of the estimand targeted by their methodology under heterogeneous treatment effects, which is similar to the weighted average of conditional LATEs considered in Huntington-Klein (2020) and Coussens and Spiess (2021). As such, their approach differs from ours as it changes the targeted estimand.

2 Framework and Proposed Estimator

We consider a data-generating process with a super-population $(Y(1), Y(0), D(1), D(0), Z, G)$, where $(Y(1), Y(0))$ are the potential outcomes when treated or not ($D = 1$ or 0), $(D(1), D(0))$ are the potential outcomes when encouraged or not ($Z = 1$ or 0), Z is the encouragement status, and G is a discrete pre-determined covariate (assumed binary in this section for illustrative purposes).[15] We have:

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$$
$$D = Z \cdot D(1) + (1 - Z) \cdot D(0)$$

Hence we consider the simple (yet common in empirical work) case where the treatment D and the instrument Z are binary. We sample n independent and identically distributed observations $\{(Y_i, D_i, Z_i, G_i)\}_{1,\dots,n}$ from this superpopulation. We work under the standard identifying assumptions of the LATE (Angrist et al., 1996) stated below.

ASSUMPTION 1 (LATE identifying assumptions).

1. *Independence:* $(Y(1), Y(0), D(1), D(0), G) \perp Z$[16]
2. *Exclusion restriction:* $Y(D, Z) = Y(D)$
3. *First Stage:* $\mathrm{E}[D = 1 \mid Z = 1] - \mathrm{E}[D = 1 \mid Z = 0] > 0$
4. *Monotonicity:* $D(1) \geq D(0)$

The only additional assumption compared to the framework considered in Angrist et al. (1996) is the independence of the covariates G and the instrument Z.[17] This is trivially satisfied for any covariates that would be determined prior to the draw of the instrument Z. Under this set of assumptions, Angrist et al. (1996) showed that the LATE, defined as the average treatment

[15] Assuming a discrete covariate is restrictive. Yet it is not uncommon in empirical work (especially when analyzing experimental data) to use discretized covariates — as it makes econometric analyses more transparent.

[16] In natural experiments, such assumption might only hold conditional on some observables. For now, we do not consider this case, and our results only apply to controlled or natural experiments that would fulfill this unconditional independence condition. Yet we conjecture that some of our results could be extended to the conditional independence case without too much additional work. We leave this for future research.

[17] In Angrist et al. (1996), the authors consider a setting where there are not any covariates in addition to Y, D and Z.

effect among compliers $\mathrm{E}[Y(1) - Y(0)|D(1) > D(0)]$, is identified. The usual estimator for the LATE is the Wald estimator — which coincides with the two-stage-least-squares (2SLS) estimator in our case where D and Z are binary:

$$\begin{aligned}\widehat{\mathrm{LATE}}^{2SLS} &= \frac{\left(\sum_i Z_i\right)^{-1}\sum_i Z_i Y_i - \left(\sum_i (1-Z_i)\right)^{-1}\sum_i (1-Z_i)\, Y_i}{\left(\sum_i Z_i\right)^{-1}\sum_i Z_i D_i - \left(\sum_i (1-Z_i)\right)^{-1}\sum_i (1-Z_i)\, D_i} \\ &= \frac{\mathrm{E}_n[Y|Z=1] - \mathrm{E}_n[Y|Z=0]}{\mathrm{E}_n[D|Z=1] - \mathrm{E}_n[D|Z=0]}\end{aligned}$$

where E_n denotes the empirical mean operator.

For illustrative purposes, consider the case where researchers have access to a binary pre-determined covariate $G \in \{0, 1\}$. By "pre-determined", we mean that G is unaffected by Z nor D — as it is determined before the realization of Z and D. To fix ideas, let us think of G as a sex indicator taking value 0 for women, 1 for men. We allow for heterogeneous shares of compliers across sex, i.e., women might react more (or less) than men to the encouragement. Formally:

$$\pi^0 \equiv \mathrm{E}\left[D(1) - D(0) \mid G = 0\right] \neq \mathrm{E}\left[D(1) - D(0) \mid G = 1\right] \equiv \pi^1.$$

We do not impose that both π^0 and π^1 are strictly larger than 0, only that the average share of compliers in the population is well separated from 0 (assumption 1.3). In other words, we allow for one of the two groups to be absolutely fully unresponsive to the encouragement — as long as the other is, allowing the identification of the overall LATE. This is key in our reasoning, as considering the existence of sub-populations with few compliers[18] (or no compliers at all) is what creates room for precision gains in the estimation of the (overall) LATE.[19] Consider the extreme case where women's share of compliers is $\pi^0 = 0$, when men's share is $\pi^1 > 0$. In such a case, women's observations do not bring any signal in the estimation of the overall LATE, as

[18] This vague terminology ("few" compliers) will be translated later in the paper in the concept of a "weak" share of compliers — i.e., a "local-to-zero" compliance rate that vanishes at a $1/\sqrt{n}$ rate (Staiger and Stock, 1997).

[19] By "overall" LATE, we mean the LATE across all groups defined by G, $\mathrm{E}[Y(1) - Y(0)|D(1) > D(0)]$, as opposed to the LATE within a given group $G = g$, $\mathrm{E}[Y(1) - Y(0)|D(1) > D(0), G = g]$. The two are by the law of iterated expectations related as follows:

$$\mathrm{E}[Y(1) - Y(0)|D(1) > D(0)] = \sum_g \mathrm{E}[Y(1) - Y(0)|D(1) > D(0), G = x] \cdot \Pr[G = g|D(1) > D(0)]$$

none of the compliers are women:

$$\begin{aligned}\text{LATE} = \; & \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), G=1]\cdot \overbrace{\mathrm{P}[G=1|D(1)>D(0)]}^{=1} \\ & + \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), G=0]\cdot \underbrace{\mathrm{P}[G=0|D(1)>D(0)]}_{=0} \\ = \; & \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), G=1]\end{aligned}$$

Neither do they prevent us from getting a consistent estimator of the LATE, as the difference in outcomes among encouraged vs. control women in the numerator of the usual LATE estimator cancels out on average (see equations below). Yet they do bring additional noise to the estimation procedure, worsening the precision of the estimator.

$$\begin{aligned}\widehat{\text{Wald}}_n &= \frac{\mathrm{E}_n[Y|Z=1]-\mathrm{E}_n[Y|Z=0]}{\mathrm{E}_n[D|Z=1]-\mathrm{E}_n[D|Z=0]} \\ &= \frac{\Delta_n^{Y|G=1}\cdot \mathrm{P}_n[G=1] + \overbrace{\Delta_n^{Y|G=0}\cdot \mathrm{P}_n[G=0]}^{\text{Mean-zero noise}}}{\mathrm{E}_n[D|Z=1]-\mathrm{E}_n[D|Z=0]}\end{aligned}$$

where $\Delta_n^{W|G=g} \equiv \mathrm{E}_n[W|Z=1, G=g] - \mathrm{E}_n[W|Z=0, G=g]$. As already mentioned in the introduction, this is easily seen when comparing the variance of a 2SLS estimator that would be computed on the sample of men only ($\mathrm{V}^{TSLS,G=1}$) with the one of a 2SLS estimator on the full sample (V^{TSLS}), assuming homoscedasticity of the errors:

$$\begin{aligned}\mathrm{V}^{TSLS} &= \frac{1}{N}\cdot\frac{1}{(\pi^1\cdot \mathrm{P}[G=1])^2}\cdot\frac{\sigma_\varepsilon^2}{p\cdot(1-p)} \\ \mathrm{V}^{TSLS,\ G=1} &= \frac{1}{N\cdot \mathrm{P}[G=1]}\cdot\frac{1}{(\pi^1)^2}\cdot\frac{\sigma_\varepsilon^2}{p\cdot(1-p)} \\ &= (1-\mathrm{Pr}[G=0])\cdot \mathrm{V}^{TSLS} < \mathrm{V}^{TSLS}\end{aligned}$$

where σ_ε^2 denotes the variance of the errors,[20] N is the sample size and $p = E[Z]$ is the share of encouraged individuals. Excluding the group without compliers ($G=0$) from the estimation

[20] Here, ε is the structural error term in what is usually called the "second stage" equation, i.e., the regression of the outcome on the treatment variable. Formally: $\varepsilon = Y - \text{LATE}\cdot D$.

decreases the variance by a factor $(1 - \Pr[G = 0])$. This is intuitive: the more we can get rid of groups without compliers, the larger the precision gains.

Motivated by this illustrative example, we propose the following estimation procedure (Estimator 1), which we will call the "naïve" Test-and-Select (naïve TS) estimator.[21]

Estimator 1 "Naïve" Test-and-Select

1: For each group defined by G: t-test on the first stage coefficient π^g. Set a given level α for the test (e.g., 5%).

2: Select only groups for which we reject the null of $\pi^g = 0$ against the alternative $\pi > 0$ (or $\pi < 0$) at a pre-specified level α (e.g., 0.05).

3: Compute the usual Wald/TSLS estimator on the selected sample.

Compared to our example, the main challenge lies in the need to pre-test on the first-stage coefficients in order to determine what are the groups without compliers. Pre-testing can create challenges for inference (Leeb and Pötscher, 2005), and recent work underlined issues with the specific procedure of pre-testing on the first-stage in IV estimation (Abadie et al., 2022). The following lemma shows that pre-testing as suggested above and estimating in the same sample will lead to a first-order bias in the estimation of the LATE parameter.

LEMMA 1 (Pre-testing and first-order bias in LATE estimation). *Let G be a binary covariate partitioning the population such that the share of compliers in groups $G = 0$ and $G = 1$ are respectively given by $\pi^0 = 0$ and $\pi^1 > 0$. Selecting groups based on a one-sided t-test with fixed test size on group-specific first-stage coefficients will lead to a first-order bias in the estimation of the LATE parameter.*

Proof. See appendix A.1. □

Lemma 1 states that there might be significant distortions due to the pre-testing step of the suggested procedure that ultimately could lead to non-valid inference. There are two sources of first-order bias introduced by this pre-testing procedure, as we make it clear in the proof of

[21] Usually, in the context of RCTs, researchers will have a strong prior about the way their encouragement affects the treatment status, hence the ability to use as an alternative hypothesis $\pi > 0$ (or $\pi < 0$) instead of $\pi \neq 0$ (see step 2 in Estimator 1). Andrews and Armstrong (2017) propose an unbiased estimator of the LATE (as an alternative to the 2SLS estimator, which is consistent yet biased in finite samples) in such cases where researchers know ex-ante the sign of the first-stage. We do not consider the use of such estimator for now in this paper.

lemma 1. The first is that this pre-test leads to an overestimation of the first-stage coefficient in the group that does not contain any compliers. This logically tends to shrink the LATE estimator (in which the overall first-stage estimator appears in the denominator) towards 0. The second source of first-order bias come from the fact that in group $G = 0$ (the one without any compliers), we end up comparing always-takers with never-takers once we condition on the estimated first-stage $\hat{\pi}^0$ being larger than a threshold. This is not an issue when the expected outcome of always takers and never-takers is the same, as this difference will concentrate around zero in this case. Yet there is no reason for these expected outcomes to coincide. When they differ, then their comparison leads to the introduction of an additional first-order bias.[22]

Simulations presented in appendix A.3 tend to confirm such concerns. We report in Table 1.1 below the results of a Monte-Carlo simulation using DGP0 described in appendix A.3. In summary, this DGP generates a sample of size $N = 1000$, divided randomly into 30 groups (i.e., roughly 33 observations per group). The share of compliers in the sample (and thus in each randomly created group on average) is 25%. In such a setting, we do not expect our procedure to yield any gains, as there are no sub-populations without compliers. Yet selecting "naively" based on a t-test — without any sample split to alleviate the pre-testing issues mentioned above — could introduce a bias in the estimation of the LATE (see lemma 1) that could invalidate the inference conducted based on such estimator. In order to provide additional evidence on this issue, Table 1.1 reports the bias and coverage rate of 95%-confidence intervals of three estimators of the LATE over 10,000 Monte-Carlo repetitions. The first column reports the performance of the 2SLS estimator, the second column the performance of our proposed methodology *with* sample splitting and cross-fitting, and the third column a "naïve" version of our methodology that would test, select and estimate the LATE in the same sample without any sample split. The results show that the naïve version of the Test-and-Select estimator exhibits a clear bias (-0.221), which is ultimately detrimental to the coverage of its associated $95\%-$confidence interval that fail to cover at their nominal rate (0.861).

[22]There is no reason for these two sources of bias to counterbalance one another. The comparison of the expected outcomes of always-takers and never-takers can either lead to a downward or upward bias on the estimated LATE, depending on whether the expected outcome of always-takers is larger (upward bias) or smaller (downward bias) than the one of never-takers.

Given the issues documented with the "naive" approach presented above, we propose a modified procedure that aims at solving the problems associated with pre-testing, building on data-splitting and cross-fitting. This Test-and-Select (TS) estimation procedure is described below (Estimator 2).

Estimator 2 Test-and-Select

1: Divide the sample in two equally sized random sub-samples $\mathcal{I}_1$ and $\mathcal{I}_2$ — stratifying the random split by G.
2: Within subsample $\mathcal{I}_1$: for each group defined by G: t-test on the first stage coefficient π^g. Set a given level α for the test (e.g., 5%).
3: Select in subsample $\mathcal{I}_2$ the groups for which we rejected — in sample $\mathcal{I}_1$ — the null of $\pi^g = 0$ against the alternative $\pi > 0$ (or $\pi < 0$) at a pre-specified level α (e.g., 0.05).
4: Compute the usual Wald/TSLS estimator on the selected sub-sample of $\mathcal{I}_2$.
5: Repeat steps 2 to 4 reversing the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ (cross-fitting).
6: Take the average of the estimators obtained in step 4 within $\mathcal{I}_1$ and $\mathcal{I}_2$.

Our proposed methodology that associates the Test-and-Select procedure with sample splitting and (2-fold) cross-fitting yields a much less biased estimator (0.097), and valid coverage (0.976) in Table 1.1. The remaining bias despite the use of data splitting and cross-fitting could be explained by the finite sample bias of 2SLS estimator.[23]

Therefore, one of the main contributions of this work is to develop valid procedures to implement the selection of groups with or without compliers in a given sample. In section 3 and as already introduced above, we propose to use data-splitting to fix the pre-testing issues previously mentioned, and we suggest the use of cross-fitting to alleviate the efficiency loss incurred when using data-splitting.

[23]Indeed, ultimately our Test-and-Select procedure with cross-fitting estimates the LATE by 2SLS on a smaller sample than the standard 2SLS estimator presented in the first column of Table 1.1. Therefore, its larger bias (0.097 vs. 0.003) could be explained by the finite sample bias of the 2SLS estimator, that vanishes as the sample size used for estimation grows.

Table 1.1 – Pre-test bias, and the use of cross-fitting

	2SLS	Test-and-Select (with 2-fold-CF)	Test-and-select (without CF)
Bias	0.003	0.097	-0.221
Coverage	0.953	0.976	0.861

Notes: This table presents the results of a simulation using the DGP0 described in section 5, with a number of groups of 30 — i.e., around 33 observations per group. In rows, we report the bias (with respect to the LATE parameter) and the coverage rate of 95%-confidence intervals. The first column reports the performance of the 2SLS estimator, the second column the performance of our proposed methodology *with* sample splitting and cross-fitting, and the third column a "naïve" version of our methodology that would test, select and estimate the LATE in the same sample without any sample split.

3 Theoretical Results

Throughout this section, we will consider a framework with two i.i.d. samples: a *test* sample (denoted $\mathcal{I}_T$) used in order to t-test on group-specific first-stage coefficients, and an *estimation* sample (denoted $\mathcal{I}_E$) used in order to compute the resulting estimator with the selection rule induced by the tests' results in $\mathcal{I}_T$. Such samples can always be constructed from a full sample of size n, by randomly splitting it with a fraction p_T (respectively $p_E = 1 - p_T$) going to sample $\mathcal{I}_T$ (respectively $\mathcal{I}_E$). We will denote by n_T (= $p_T \cdot n$) and n_E (= $p_E \cdot n$) the corresponding sample sizes — and we will use the notation $n \to \infty$ to describe an asymptotic in both n_E and n_T simultaneously. At the end of the section, we will consider the use of cross-fitting — i.e., reversing the roles of $\mathcal{I}_T$ and $\mathcal{I}_E$ to get two estimators subsequently averaged — as an attempt to mitigate the loss of precision induced by sample splitting.

The study of the properties of our suggested estimator will be divided into two parts. Firstly, we will consider the case where covariate-defined sub-groups contain either a share of compliers well-separated from zero, or no compliers at all. This case will simplify the study of the potential precision gains derived from the suggested procedure. In a second step, we will introduce groups with a "local-to-zero" (or "weak") share of compliers, à la Staiger and Stock (1997) — meaning that the share of compliers in those groups decreases at a $1/\sqrt{n}$ rate, placing them in the same order of magnitude as sampling variation. Such a modeling choice is made in an effort to better approximate the finite-sample behavior of the estimator, by allowing for imperfect selection of

groups with non-zero shares of compliers.[24] Recall from the previous section introducing our framework that our population is partitioned by a grouping variable G. Following the notations introduced in this previous section, we will denote by π^g the share of compliers in group $G = g$. We denote by $\mathcal{G}$ the support of G. In order to distinguish groups with "strong", "weak" and zero shares of compliers, we will further define:

1. $\mathcal{G}_S = \{\text{all groups with strong first stage}\}$
2. $\mathcal{G}_W = \{\text{all groups with weak first stage}\}$
3. $\mathcal{G}_0 = \{\text{all groups with zero first stage}\}$

3.1 Standard asymptotics

In this section, we will work under the assumption that there are only two types of groups: the ones without any compliers, and the ones with a strong first-stage (i.e., a share of compliers well separated from 0).

ASSUMPTION 2 (No weak first-stages). *There are no groups for which the share of compliers is local-to-zero. Formally: $\mathcal{G}_W = \varnothing$.*

Let $S \in \{0,1\}^{|\mathcal{G}|}$ denote an arbitrary selection vector, where $S_g = 1$ indicates that group $G = g$ is selected in the restricted sample used for estimation in our proposed procedure. Let us define the selected estimator:

$$\begin{aligned}\hat{\tau}(S) &= \frac{\left(\sum_{i|S_{G_i}=1} Z_i\right)^{-1} \sum_{i|S_{G_i}=1} Z_i Y_i - \left(\sum_{i|S_{G_i}=1} (1-Z_i)\right)^{-1} \sum_{i|S_{G_i}=1} (1-Z_i)\, Y_i}{\left(\sum_{i|S_{G_i}=1} Z_i\right)^{-1} \sum_{i|S_{G_i}=1} Z_i D_i - \left(\sum_{i|S_{G_i}=1} (1-Z_i)\right)^{-1} \sum_{i|S_{G_i}=1} (1-Z_i)\, D_i} \\ &= \frac{\mathrm{E}_n[Y|Z=1,S_G=1] - \mathrm{E}_n[Y|Z=0,S_G=1]}{\mathrm{E}_n[D|Z=1,S_G=1] - \mathrm{E}_n[D|Z=0,S_G=1]}\end{aligned}$$

In words, $\hat{\tau}(S)$ is the Wald estimator on the subsample such that $S_{G_i} = 1$, which is the subsample

[24] An alternative modeling choice would consider a growing number of groups, so that the number of observations per group could remain stable as the overall sample size goes to infinity. This is not our framework here: the share that each group g represents in the population is assumed stable with respect to the sample size. We shall investigate in future versions of this work whether this alternative modeling brings new insights.

designated by S. As an example, for $|\mathcal{G}| = 2$,

$$\hat{\tau}(S) = S_1 S_0 \hat{\tau}^{WALD} + S_1(1 - S_0)\hat{\tau}_1^{WALD} + S_0(1 - S_1)\hat{\tau}_0^{WALD} \tag{1.1}$$

$$= \begin{cases} \hat{\tau}^{WALD} & \text{if} \quad S_1 = S_0 = 1 \\ \hat{\tau}_1^{WALD} & \text{if} \quad S_1 = 1 \quad \& \quad S_0 = 0 \\ \hat{\tau}_0^{WALD} & \text{if} \quad S_1 = 0 \quad \& \quad S_0 = 1 \end{cases} \tag{1.2}$$

where $\hat{\tau}_g^{WALD}$ denotes the Wald estimator computed on the observations with $G = g$. The selection vector S of interest here is the one determined through group-by-group t-tests in the test sample $\mathcal{I}_T$ — constructed as a random split of the initial sample.[25] We will denote the latter by $\hat{S}^{(T)}$, where the hat and superscript (T) indicate that this vector comes from an estimation step in sample $\mathcal{I}_T$. We can then define:

$$\hat{\tau}_E = \hat{\tau}\left(\hat{S}_{(T)}\right) \tag{1.3}$$

which is ultimately our estimator of interest, the TS estimator computed on split $\mathcal{I}_E$.[26] Equivalently, for any selection vector S, we will denote by $\hat{\tau}_E(S)$ the estimator computed on the subsample defined by S, within split $\mathcal{I}_E$.

We start by characterizing the asymptotic behavior of this selection procedure.

LEMMA 2 (Asymptotic distribution of the selection procedure). *Under assumptions 1 and 2, and if $E[|Y|^2] < \infty$, as the test sample size n_T goes to infinity, the probability of selecting groups with a first stage of 0 goes to α (the level of the t-test used) and the probability of selecting groups with strong first-stages goes to 1.*

Proof. See appendix A.2.

□

Notice that it would be possible to decrease the threshold of the t-test at an appropriate rate

[25] This vector stacks the $|\mathcal{G}|$ test decisions resulting from our $|\mathcal{G}|$ t-tests (one per group) in $\mathcal{I}_T$.

[26] We consider the use of cross-fitting, leading to the use of the "symmetric" estimator $\hat{\tau}_T = \hat{\tau}\left(\hat{S}_{(E)}\right)$ later in this section.

so that the probability to exclude groups with no first-stages goes to 1 as the sample size goes to infinity. Yet the resulting asymptotic approximation would likely not reflect accurately what happens in finite samples — in which the likelihood of keeping groups with zero first-stages would remain positive — hence we do not consider such type of testing for our selection procedure. Lemma 2 shows that groups with strong first stages will always be selected asymptotically. Hence, when studying the asymptotic distribution of $\hat{\tau}_E(S)$ when both the test and estimation sample sizes (n_T and n_E) tends to infinity, we can restrict ourselves to vectors S which select at least all vectors with strong first stages. We denote by $\mathcal{S}_{strong}$ this subset of all selection vectors (i.e., a subset of $\{0,1\}^{|\mathcal{G}|}$) that never excludes groups with strong first-stages. Formally, for any $\tilde{S} \in \mathcal{S}_{strong}$, we have: $\forall g \in \mathcal{G}_S,\ \tilde{S}_g = 1$.

PROPOSITION 1. *Under assumptions 1 and 2, and if $E[Y^2] < \infty$, then we have:*

1. $\forall S \in \mathcal{S}_{strong}$, $\sqrt{n_E}\left(\hat{\tau}_E(S) - LATE\right) \xrightarrow{d} \mathcal{N}(0, V^{\hat{\tau}_E(S)})$
2. $\forall S \in \mathcal{S}_{strong}$, $V^{\hat{\tau}_E(S)} \leq V^{TSLS}$ *with equality iff:* $\forall g,\quad S_g = 1$ *or in degenerate cases*
3. *We have:*
$$\sqrt{n_E} \cdot \frac{\hat{\tau}_E - LATE}{\sqrt{V^{\hat{\tau}_E}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

Proof. See appendix A.1. □

For any realization of $\hat{S}$ denoted $S \in \mathcal{S}_{strong}$, one can build asymptotically valid confidence intervals with coverage $(1-\alpha)$ *conditional* on the realization of $\hat{S}$ in the usual way:

$$CI_\alpha(S) = \left[\hat{\tau}_E(S) - \frac{\sqrt{\hat{V}^{\hat{\tau}_E(S)}}}{\sqrt{n_E}} \cdot q_{1-\frac{\alpha}{2}}\ ,\ \hat{\tau}_E(S) + \frac{\sqrt{\hat{V}^{\hat{\tau}_E(S)}}}{\sqrt{n_E}} \cdot q_{1-\frac{\alpha}{2}}\right]$$

where $\hat{V}^{\hat{\tau}_E(S)}$ is a consistent estimator of the asymptotic variance of $\hat{\tau}_E(S)$, and $q_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the $\mathcal{N}(0,1)$ distribution. Those CIs are asymptotically valid by proposition 1.1, i.e.:

$$\mathrm{P}[LATE \in CI_\alpha(S)] \xrightarrow[n\to\infty]{} 1-\alpha$$

The following corollary states that such intervals have asymptotically valid *unconditional* coverage for the LATE. It also states that when the selection S is such that the asymptotic variance of

the resulting estimator is strictly lower than the one of the TSLS estimator (inequality case of proposition 1.2), then the length of a CI conditional on such an S is going to be lower than usual CIs based on the TSLS estimator with probability going to 1 as n goes to infinity — reflecting the gains in terms of inference. Notice that the asymptotic study of CIs lengths requires rescaling CIs by $\sqrt{n_E}$ to allow for a meaningful comparison.[27]

COROLLARY 1. *Under assumptions 1 and 2, if $E[Y^2] < \infty$ and S is such that we are in the inequality case of proposition 1.2, then the estimators $\hat{\tau}_E(S)$ and $\hat{\tau}_E^{TSLS}$ (the TS estimator conditional on S and TSLS estimator computed in split $\mathcal{I}_E$) are such that:*

$$\lim_{n\to\infty} \mathrm{P}\left[\sqrt{n_E} \cdot \text{length}[CI_\alpha(S)] \leq \sqrt{n_E} \cdot \text{length}[CI_\alpha^{TSLS}]\right] \;=\; 1$$

Moreover, we have that:

$$\mathrm{P}\left[LATE \in CI_\alpha(\hat{S})\right] \underset{n\to\infty}{\longrightarrow} 1-\alpha$$

where $\hat{S}$ is the (random) selection vector estimated from the test data $\mathcal{I}_T$.

Proof. See appendix A.1.

□

Proposition 1 and corollary 1 show — under assumption 2 ruling out the presence of sub-populations with weak first-stages — that our procedure dominates unequivocally the usual approach (based on TSLS/Wald estimator) for estimation of and inference on the LATE parameter. Yet it should be noted that the use of sample splitting was key to derive those results, as it allowed us to consider as independent the selection process and the estimation. And the variance comparison made in proposition 1.2 between our TS procedure and the 2SLS estimator is based on a comparison of asymptotic variances, while the second statement of corollary 1 assumes that the sample size used for estimation are identical when implementing our TS strategy or the usual TSLS estimation approach. But given the sample splitting step inherent to our methodology, a fair comparison between the inference derived from the TSLS approach and our proposed

[27] Otherwise, any CI constructed in the usual way based on asymptotically normal estimators for a point-identified parameter will have a length that shrinks to 0 (at a $\sqrt{n_E}$ rate).

strategy should take into account the reduction in sample size in the latter approach. Indeed, this reduced sample size tempers the gains in asymptotic variance. A simple numeric example inspired from the one presented in the introduction illustrates this issue. Suppose again an experiment with a 10% compliance rate in the whole population, yet where compliers are all concentrated in a sub-population representing half of the total population. In principle, if the researcher had some additional pilot sample allowing her to test and restrict accordingly the estimation to this compliant population, then the variance of the estimator would be halved — compared to the variance of the usual TSLS estimator, see equation 1.[28] Indeed, the sample size used for estimation is divided by 2 (doubling the variance of the estimator all else equal), yet the compliance rate is doubled, dividing by 4 the variance. Yet in general, the researcher won't have an additional separate sample to implement the testing step. In this case, she will need to (randomly) split her sample in two sub-samples to implement our methodology, reducing the size of the estimation sample in comparison to the usual TSLS estimation. Suppose she implements a 20%-80% split to create a test and estimation sample.[29] Then instead of dividing by two the size of the estimation sample (post-selection), she ends up reducing it by a factor of $\frac{4}{5} \cdot \frac{1}{2} = \frac{2}{5} < \frac{1}{2}$ — compared to the sample size used in TSLS estimation. Hence the reduction in variance goes from a factor of $\frac{1}{2}$ to a factor of $\frac{5}{2} \cdot \frac{1}{4} = \frac{5}{8} > \frac{1}{2}$. More generally, if the gains in variance derived from the increased compliance rate in the selected population aren't large enough, they can be cancelled by the losses due to the sample split — to the point that the overall procedure might lead to an *inflated* variance in the worst cases.

Cross-fitting The example above makes it clear that the sample splitting step is not innocuous for precision, due to the loss in the sample size effectively used in the estimation step. Yet it is a key step of our approach as it allows making the testing-selecting and estimation steps independent. As shown in lemma 1 and illustrated in Table 1.1, our procedure would yield a biased estimator in the absence of sample splitting.

[28] This is assuming homoscedasticity in order to simplify the computations for illustrative purposes.

[29] There isn't a clear way to determine the proper splitting rule between a test and estimation sample. In principle, the test sample only needs to be large enough so that asymptotic approximations *within each group* are valid. The remaining of the initial sample should be assigned to the estimation step, as the purpose of this strategy is ultimately to improve inference.

Ideally, one would like to benefit from the advantages of sample splitting without facing the precision loss due to burning a fraction of the sample in the testing-selecting step. A way to do so consists in using both splits of the sample for both the testing-selecting and estimation steps by reversing their roles — what is usually called cross-fitting in the machine learning literature. In other words, the researcher divides the sample in two (or more) equally-sized folds, $\mathcal{I}_1$ and $\mathcal{I}_2$. She constructs a first estimator using $\mathcal{I}_1$ as the test sample and $\mathcal{I}_2$ as the estimation sample, and a second using $\mathcal{I}_2$ as the test sample and $\mathcal{I}_1$ as the estimation sample (see the description of our procedure in section 2, Estimator 2). This way, all the sample is used for estimation, and the hope to recover some form of efficiency is revived. Indeed, the two (or more, if more folds are created) estimators constructed in this way benefit from the same gains in (asymptotic) variance than the ones discussed above for the sample split estimator. Hence averaging those estimators would potentially yield an estimator with the same variance as a hypothetical one constructed using the full sample, with an additional independent test sample used for selection. A sufficient condition for such gains in variance to be restored is that the two cross-fit estimators are independent one from another. This is what the following lemma establishes.

LEMMA 3 (Independence of cross-fit estimators). *Under assumptions 1 and 2, two estimators constructed following our suggested procedure and reversing the roles of two independent samples $\mathcal{I}_1$ and $\mathcal{I}_2$ are asymptotically independent one from another.*

Proof. See appendix A.1. □

Cross-fitting is therefore a way to restore the full variance gains described in the above section, despite the use of sample splitting. Indeed, the asymptotic variance of the average of $\hat{\tau}_1$ and $\hat{\tau}_2$ is given by:

$$V\left(\frac{\mathcal{N}(0, V^{\hat{\tau}_1}) + \mathcal{N}(0, V^{\hat{\tau}_2})}{2}\right) = \frac{V^{\hat{\tau}_1} + V^{\hat{\tau}_2}}{4} = \frac{V^{\hat{\tau}_1}}{2}$$

where the first equality uses the independence between the limiting distributions of $\hat{\tau}_1$ and $\hat{\tau}_2$ demonstrated in lemma 3. Hence our cross-fitted TS estimator $(\hat{\tau}_1 + \hat{\tau}_2)/2$ has an asymptotic variance that is half the one of an estimator computed on a single split. In parallel, the sample splitting step results in a loss of a factor $\sqrt{2}$ in the speed of convergence (compared to the speed

of convergence of an hypothetical TS estimator that could be computed on the whole sample of size n). Therefore, the gain in asymptotic variance described in above exactly compensates the precision loss due to the sample split.

The above results are encouraging as they suggest that *asymptotically* there are indeed gains in precision from testing and selecting a sub-sample with statistically significant first-stages. Yet as already vastly documented in the statistical and econometrics literature, pre-testing methods should be treated with caution as standard asymptotic approximations of these procedures can often be misleading.[30] In particular, our framework so far ruled out the possibility to wrongly exclude groups with some compliers — as by consistency of the t-test against any (well-separated from 0) alternative, the probability to exclude such groups from the selected sample was asymptotically zero. This is not a satisfactory approximation of what would happen in finite samples — in which groups with small shares of compliers might be wrongly excluded by the selection procedure. Therefore, we need to extend our framework in order to account for such cases.

3.2 Asymptotic results with "weak" first-stages

Now, we introduce groups with local-to-zero first-stages. Those groups are such that their share of compliers evolves at the same rate as $1/\sqrt{n}$, so that a t-test will not systematically conclude that the first-stage coefficient is different from zero even with a sample size going to infinity.

ASSUMPTION 3 (Weak first-stages, fixed shares and fixed conditional LATEs). *There are groups with a local-to-zero share of compliers. Formally:*

$$\exists g \in \mathcal{G} \text{ s.t. } \pi_n^g = \frac{H^g}{\sqrt{n}}, \text{ with } H^g \in \mathbb{R}^+ \backslash \{0\}$$

All such values of g for which first-stages are weak are gathered in $\mathcal{G}_W$.

In parallel, the data-generating process is assumed to be such that for any group g, the share of observations

[30] For a seminal exposition to these issues, see Leeb and Pötscher (2005).

contained in the group is constant (it does not vary with n), nor does the LATE within the group. Formally:

$$\forall g \in \mathcal{G},\ \forall n,\ \mathrm{P}[G = g] = p_g \in (0, 1)$$
$$E[Y(1) - Y(0) | D(1) > D(0),\ G = g] = l_g \in \mathbb{R}$$

One should still keep in mind that we maintain the assumption of a strong first-stage overall (see assumption 1), meaning that:

$$\forall n, \pi = \sum_{g=1}^{|\mathcal{G}|} \pi_n^g \geq c > 0$$

where c is a constant that does not depend on n. In other words, we still assume that there are some groups with strong first-stages in the population. Moving away from such a setting would place ourselves in the realm of weak-identification, which is not the focus of our work here. Instead, we consider settings in which identification strength is high enough, and precision of the estimation procedure is the "only" problem to be fixed (if and when possible).

We start by characterizing the asymptotic behavior of the selection procedure when there are some groups with weak first stages.

LEMMA 4 (Asymptotic distribution of the selection procedure with some weak group first-stages). *Under assumptions 1 and 3, and if $E[|Y^{2+\delta}|] < \infty$ (for some $\delta > 0$), as the test sample size n_T goes to infinity, the probability to select groups with 0 first stages goes to α (the level of the t-test used), the probability to select groups with strong first-stages goes to 1, and the probability to select groups with weak ("local-to-zero") first-stages goes to values in the $[\alpha,\ 1)$ range — depending on the localization parameter H^g.*

Proof. See appendix A.2. □

As in lemma 2, lemma 4 above justifies that when studying the asymptotic distribution of $\hat{\tau}(S)$ as both the test and estimation sample size tend to infinity, we only consider selection vectors S that satisfy: $\forall g \in \mathcal{G}_S,\ S_g = 1$ (where S_g denotes the g-th term of vector S). This is because asymptotically, we won't make any exclusion error regarding groups with strong first-stages,

that will always be selected in the estimation sample. Yet this is not the case for groups with weak first-stages, as we will exclude them with a non-zero probability (even asymptotically) despite their non-zero share of compliers. In the previous subsection 3.1 and its associated proposition 1, we showed that in the absence of such groups with weak first-stages, our estimator could yield precision gains without introducing any first-order bias. The following proposition (the analog to proposition 1) shows that it is no longer true in the presence of some weak first-stages.

PROPOSITION 2. *Under assumptions 1 and 3, and if* $E[|Y^{2+\delta}|] < \infty$ *(for some* $\delta > 0$*), we have:*

1. $\forall S \in \mathcal{S}_{strong}$,[31] $\quad \sqrt{n_E}\,(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(B(S), V^S))$.
2. $B(S) \propto |LATE^{W(S)} - LATE|$, *where* $LATE^{W(S)}$ *denotes the average treatment effect among compliers within groups with weak first-stages that are wrongly dropped by selection procedure* S.
3. $B(S) \neq 0$ *if* $\exists j \; s.t. \; \{S_j = 0 \cap j \in \mathcal{G}_W\}$ *and* $LATE^{W(S)} \neq LATE$.

Proof. See appendix A.1. □

Without any further assumptions on treatment effect heterogeneity, the above proposition suggests that our proposed estimator will systematically be first-order biased in the presence of groups with weak first-stages. Indeed, the probability of wrongly excluding those groups does not go to zero asymptotically (see lemma 4) and proposition 2.3 shows that in the presence of such exclusion errors, the first-order bias of our procedure is non-zero. The intuition behind such a bias is relatively simple: the LATE within groups that contain a weak share of compliers might differ from the LATE within groups that are kept for the estimation step. If we were to bundle all groups with a weak first-stage in a single group $G = 2$, and all groups with a strong first-stage in $G = 1$, the asymptotic bias (conditional on the event that group $G = 2$ is dropped from the estimation step) would take the following form:

$$B = \underbrace{\frac{H^2 \cdot \Pr[G = 2]}{\pi}}_{\substack{\text{Sh. of compliers w/ G=2} \\ \text{among all compliers}(\times\sqrt{n})}} \cdot \underbrace{\left(LATE^1 - LATE^2\right)}_{\substack{\text{Treatment effect} \\ \text{heterogeneity}}}$$

where $\pi \equiv \Pr[D(1) > D(0)]$ is the share of compliers in the population, $LATE^g \equiv E[Y(1) -$

[31] Recall that $\mathcal{S}_{strong}$ is defined such that for any $\tilde{S} \in \mathcal{S}_{strong}$, we have: $\forall g \in \mathcal{G}_S, \; \tilde{S}_g = 1$.

$Y(0) \mid D(1) > D(0),\ G = g]$ is the LATE in group $G = g$ and H^2 is the localization parameter for the first stage in group $G = 2$. The reason why this is "only" a first-order bias can also be seen in the above display. Indeed, the share of compliers with $G = 2$ among all compliers decreases at a $\sqrt{n}$-rate under assumption 3. Hence even once rescaled by $\sqrt{n}$, the bias (with respect to the LATE parameter) remains bounded as long as the treatment effect heterogeneity term $(LATE^1 - LATE^2)$ is bounded.

In order to better grasp the nature of the first-order bias of our estimator, corollary 2 provides sufficient conditions on treatment effect heterogeneity for our estimator to remain first-order unbiased.

COROLLARY 2. *Under assumptions 1 and 3, $E[|Y^{2+\delta}|] < \infty$ (for some $\delta > 0$), and homogeneous treatment effects, we have that $\hat{\tau}(S)$ is first-order unbiased and asymptotically normal, i.e.:*

$$\forall S \in \mathcal{S}_{strong}, \quad \sqrt{n_E}\,(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(0, V^S)).$$

Less restrictively, under assumptions 1 and 3, $E[|Y^{2+\delta}|] < \infty$ (for some $\delta > 0$), and vanishing treatment effect heterogeneity, i.e.:

$$\forall g \in \mathcal{G}_W, \quad |LATE_g - LATE| = o(1)$$

$\hat{\tau}(S)$ is also first-order unbiased and asymptotically normal.

Assuming homogeneous treatment effect is not realistic either, and rather in opposition to the spirit of the LATE literature. On the other hand, vanishing treatment heterogeneity might be a realistic assumption to describe the data-generating processes studied in applied economics and social sciences in general. For instance, Coussens and Spiess (2021) studied the properties of their proposed estimator under the assumption that treatment effect heterogeneity would be of the same order of magnitude as sampling variation, i.e., decreasing at a $1/\sqrt{n}$ rate. This type of restriction can be motivated by the usual difficulties faced by researchers in detecting treatment effect heterogeneity in empirical research, given the usual sample sizes at their disposal. Let us consider what are the properties of our estimator under such restrictions placed on treatment

effect heterogeneity.[32]

ASSUMPTION 4 (First order negligible heterogeneity or noisy heterogeneity). *The heterogeneity of conditional LATEs across groups is of the same order of magnitude as the sampling variation. Formally:*

$$\forall g \in \mathcal{G}_W, \quad |LATE_g - LATE| = O(n^{-1/2})$$

The next theorem studies the asymptotic distribution of our estimator in such a framework. Notice that the results presented in the theorem below would hold under the less stringent assumption of vanishing treatment effect heterogeneity, i.e.:

$$\forall g \in \mathcal{G}_W, \quad |LATE_g - LATE| = o(1)$$

instead of assumption 4. We state it under assumption 4 in the hope that relating treatment effect heterogeneity to the order of magnitude of sampling variation would be more interpretable.

THEOREM 1. *Under assumptions 1, 3, 4, and if $E[|Y^{2+\delta}|] < \infty$ (for some $\delta > 0$), we have:*

1. $\forall S \in \mathcal{S}_{strong}$, $\sqrt{n_E}\,(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(0, V(\hat{\tau}(S)))$ *with* $V(\hat{\tau}(S)) \leq V^{TSLS}$.
2. *We have*
$$\sqrt{n_E} \cdot \frac{\hat{\tau}_E - LATE}{\sqrt{V^{\hat{\tau}_E}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

Proof. See appendix A.1. □

Theorem 1 above establishes the $\sqrt{n}$-convergence of our estimator under assumptions 3 and 4. The gains in inference already studied in the absence of any weak first-stage groups (see corollary 1) remain following the same reasoning. Compared to alternatives such as the one suggested in Coussens and Spiess (2021) — equivalent to the estimator studied in Huntington-Klein (2020) — our procedure presents the benefit of being exempt of any first-order bias under the restriction on treatment effect heterogeneity made in assumption 4 — see lemma 9 and its proof in appendix

[32] Let us note that contrary to Coussens and Spiess (2021), we do not assume that the average treatment effect in general is of the order of magnitude of $1/\sqrt{n}$, but rather that treatment effect *heterogeneity* is. We justify this further below. This seems a less stringent assumption, and is sufficient for our purposes.

A.2 for a proof of the bias of Coussens and Spiess (2021) procedure under our framework.[33] The intuition behind the relatively good behavior of our estimator can be given as follows. In the absence of any restrictions on treatment effect heterogeneity, both our estimator and the one studied by Coussens and Spiess (2021) converge to weighted averages of conditional LATEs. Yet the estimand towards which Coussens and Spiess (2021) estimator converges weights each $LATE^g$ by the square of the share of compliers in group g, creating possibly large deviations from the usual LATE parameter — which weights each $LATE^g$ by the share of compliers (not squared). Therefore, assuming that the heterogeneity across conditional LATEs is of the order of $1/\sqrt{n}$ is not sufficient to compensate for the deviations from the LATE created by the weighting scheme. On the contrary, our estimator's bias in the absence of assumption 4 is due to the failure to systematically select groups with weak shares of compliers. Hence the conditional LATEs of such groups end up being weighted less than they should to match with the overall LATE parameter. Yet for our estimator, this only affects groups with very low compliance rates, that do not represent a very large share of the total population of compliers. Hence the deviation from the LATE in our case is less important than in Coussens and Spiess (2021), and the restriction on heterogeneity made in assumption 4 is sufficient to rule out any first-order bias. We view such a discrepancy in the behavior of our estimator compared to the one of Coussens and Spiess (2021) as revealing two points:

1. the heterogeneity restriction made in assumption 4 is far from being equivalent to homogeneous treatment effects, as estimators such as the one of Coussens and Spiess (2021) that would converge to the LATE in the homogeneous case exhibit a first-order bias under this assumption ;
2. our estimator offers gains in variance while remaining more tightly related to the LATE parameter than the one studied in Coussens and Spiess (2021). Hence we offer another alternative in the bias-variance trade-off, from no asymptotic bias (yet larger variance) when using TSLS to potentially larger gains in variance when using Coussens and Spiess

[33] Coussens and Spiess (2021) already establish the bias of the estimator they study under the assumption that all conditional LATEs are local to zero. In lemma 4, we simply prove that their result still holds under our own assumption that only restricts treatment effect *heterogeneity* to be local to zero.

(2021) (at the cost of a larger asymptotic bias, even under restrictions on treatment effect heterogeneity).

Yet empirical researchers might view assumption 4 as merely a convenient theoretical device without any ground in empirical practice. We would like to offer some heuristic suggesting that such an assumption might be justified in empirically relevant settings. Consider the case of researchers implementing an encouragement design to study the impact of a given policy (e.g., training programs). A common practice is to choose the sample size to be able to detect a given magnitude of effect $\kappa\%$ of the time (where $\kappa = 80$ is the usual choice). This "minimum detectable effect" (MDE, often denoted e^*) sometimes coincides with what researchers deem to be an economically significant effect, and/or the magnitude of effects typically measured in the literature. The usual formula to express this e^* as a function of the sample size if the following:

$$e^* = \sqrt{\frac{\sigma^2}{n \cdot \mathrm{E}[Z] \cdot (1 - \mathrm{E}[Z])}} \cdot \frac{1}{\mathrm{E}(D \mid Z = 1) - \mathrm{E}(D \mid Z = 0)} \cdot \left(q_{1-\frac{\alpha}{2}} + q_\kappa\right)$$

where we assumed $\mathrm{Var}[Y(0)] = \mathrm{Var}[Y(1)] = \sigma^2$,[34] and q_x is the x^{th}-quantile of a $\mathcal{N}(0, 1)$. Hence in studies designed based on power analyses, we have by design: $e^* = O(n^{-1/2})$. It can still be that the true effect (and treatment effect heterogeneity) is way larger than e^*, in which case our study will systematically detect the effect of the policy (and its heterogeneity). This would be the case in general in sciences that are over-powered... yet social sciences (and economics in particular) have rather been documented to be *under*-powered in meta-analyses — e.g., in Ioannidis et al. (2017). Experimenters in social sciences certainly do not detect 100% of the time significant effects (and even less often treatment effect *heterogeneity*). Hence it might seem reasonable to assume that most of the time, the true effects (and true heterogeneity) is of the same order of magnitude as the MDE of the study designed to detect them. In such a case, assumption 4 would be fulfilled.

[34] I.e., under the simplifying assumption of constant (or uncorrelated with X) treatment effects, and homoscedastic errors.

4 Extensions

In this section, we present some possible extensions.

High-dimensional groups Assuming Xs can define groups with weak/0 share of compliers is arguably more credible when Xs are high dimensional (e.g., when there is a large number of covariates, interactions between covariates, continuous covariates etc.). The question then becomes: how to adapt our procedure to this setting? We will have to maintain the assumption of strong identification overall, i.e. $\pi > 0$. Then the most natural way to proceed seems to follow a strategy along the lines of Chernozhukov et al. (2021), e.g.:

1. Build a flexible prediction of $s(X) \equiv \mathrm{E}[D(1) - D(0)|X]$, denoted $\hat{s}(X)$
2. No assumption on the rate of convergence of $\hat{s}(X)$. The hope is merely that $\hat{s}(X)$ contains some signal for the true $s(X)$.
3. Define $\bar{G}$ (a fixed number) groups based on quantiles of $\hat{s}(X)$, and use Chernozhukov et al. (2021) results to make inference on:
   $\mathrm{E}[s(X)|\hat{s}(X) \in [q_{g-1}, q_g]] = \mathrm{E}[D(1) - D(0)|\hat{s}(X) \in [q_{g-1}, q_g]] \equiv \pi^g$

That way, we are back to a situation in which the covariates are reduced to a partition of the population: $\{\mathbb{I}_{\hat{s}(x)\in[q_{g-1},q_g]}\}_{g\in\{1,\dots,G\}}$. Unfortunately, the procedure proposed in Chernozhukov et al. (2021) cannot be directly applied to our setting since it is based on repeated data-splits, with $\hat{s}(X)$ being estimated repeatedly, such that inference on so-called GATEs — grouped average treatment effects, of the form $\mathrm{E}[s(X)|\hat{s}(X) \in [q_{g-1}, q_g]]$ — can be made without offering a clear way to associate a given observation to a given group — since such a mapping will change from one data-split to another. A simple fix to this issue is to commit to a single data-split. This is the choice we make in our application in section 6. Chernozhukov et al. (2021) defend instead the variational inference approach they develop, as (i) it limits the risk of p-hacking from researchers if they do not commit (e.g., by setting a seed) to a single random split and search for a "favorable" one and (ii) such commitment would expose researchers to the risk of drawing a "bad" split.

That said, the variational inference approach cannot be readily applied to our setting,[35] and as of today we have not found any alternatives to such a "commitment" in the high dimensional covariates setting.

Re-weighting strategy Instead of taking a binary decision to either drop or include groups in the estimation sample, an alternative might be to re-weight groups based on their probability to have a 0 share of compliers. This probability is directly given by the p-value associated to the t-test we were using so far for the selection decision. It is possible that such an alternative procedure could be properly motivated by a model-selection framework in which we optimally trade-off bias and variance (to minimize RMSE) by taking weighted averages of LATE estimators estimated on the full sample — lower bias, higher variance — or on a sample selected based on group-specific first-stage coefficients — higher bias, lower variance — in the spirit of Claeskens and Hjort (2003) and Kitagawa and Muris (2016). Our main results might still hold for such weighted estimator since (asymptotically) groups with strong first-stages would have $\Pr[\text{sh. of compliers} = 0 | g \in \mathcal{G}_S]$ that goes to 0, hence a weight that goes to 1 as is already the case in our proposed estimation strategy.

Notice that this would still be distinct from Coussens and Spiess (2021) "weighted-IV" approach, as our weights would tend to 1 and be uniform among all groups with a strong first-stage. This way, we could still hope that changes in the targeted estimand remain negligible under restrictions on treatment heterogeneity of the type described in assumption 4 — which is not the case for the "weighted-IV" approach (see lemma 9).

Breakdown analysis Instead of relying on an assumption of the type of assumption 4, researchers might prefer to acknowledge the potential (first-order) difference in the estimand targeted by our estimator and the LATE, and make use of sensitivity analyses to determine under which conditions some inferential statements derived based on our proposed estimator — e.g., the LATE is higher than a given threshold — might be erroneous.

[35] Indeed, this approach relies on repeating the data splitting step a certain number of times (taking median of p-values or CIs bounds at level $\alpha/2$ to construct p-values and CIs of level α. Yet in our case, repeating the splitting step would prevent us from creating a single fixed partition of the population to be used as our covariate G.

Here is one way to approach such sensitivity analyses. It relies on the observation that the gap between the estimand targeted by our procedure — the average effect among compliers within the selected population — and the original LATE (on the whole population) has the following expression:

$$B = \mathrm{P}[G = 2|D(1) > D(0)] \cdot \left(LATE^2 - LATE^1\right)$$

where $G = 2$ denotes the population not selected, $G = 1$ the selected population, and $LATE^1$and $LATE^2$ the LATEs within those two populations — i.e., $LATE^1$ denoted the estimand targeted by our procedure. Of course, $G = 1$ and $G = 2$ depend on the realization of the sample. Let us consider a sensitivity analysis that would condition on the sample realization, so that $G = 1$ and $G = 2$ are considered as deterministic.[36] The quantity $\mathrm{P}[G = 2|D(1) > D(0)]$ can be estimated by 2SLS as suggested in Abadie (2003).[37] Yet by construction of $G = 2$, Z is a weak instrument for D in this subpopulation, thus $\mathrm{P}[G = 2|D(1) > D(0)]$ cannot be consistently estimated. However, that does not prevent us from constructing an asymptotically valid $(1 - \alpha)$-confidence interval around this parameter — e.g., using inversion of an Anderson-Rubin statistic. Suppose we construct 99%-CI around $\mathrm{P}[G = 2|D(1) > D(0)]$, and take the upper bound of this quantity, denoted $\widehat{UB}^P$. The bias term B is increasing in $\mathrm{P}[G = 2|D(1) > D(0)]$, hence the *worst-case* bias can be obtained by replacing $\mathrm{P}[G = 2|D(1) > D(0)]$ with $\hat{UB}^P$. We are left with the unknown $(LATE^2 - LATE^1) \equiv M$, that is going to be our sensitivity parameter. For a given value of M, the worst-case bias of our proposed estimator for the LATE is given by $M \cdot \widehat{UB}^P$. Suppose we widen our 96%-CI around $LATE^1$ — the effect among compliers in the population selected by our procedure — by $\pm M \cdot \widehat{UB}^P$. Such CI is (asymptotically) valid with a 95% coverage for the overall LATE parameter.[38] The "breakdown" analysis would then consist in determining for which value of M the CI constructed in such a way includes a threshold value (e.g., 0). If

[36] In other words, $LATE^1$ and $LATE^2$ become estimands that are sample-dependent. This is not an issue as ultimately, this sensitivity analysis will still be related to an estimand that is sample-independent, namely the LATE in the whole population.

[37] It suffices to regress $\mathbb{1}\{G = 2\} \times D$ on D instrumented by Z.

[38] Indeed, our worst-case bias estimate is only valid with probability 0.99, as it is based on the upper bound of a 99%-CI on $\mathrm{P}[G = 2|D(1) > D(0)]$. Therefore, using 96%-CI around $LATE^1$, we get a CI that has coverage equal to $0.99 \times 0.96 = 0.9504$.

this value is very high, the analysis — and inferential statements on the LATE — based on our proposed estimation strategy could be considered robust to treatment effect heterogeneity.

5 Simulations

This section presents a simulation study that compares the performance of the various estimators mentioned above: the standard 2SLS, our proposed Test-and-Select estimator, Huntington-Klein (2020)'s interacted IV estimator and Coussens and Spiess (2021)'s compliance-weighted IV estimator. We consider a number of Data Generating Processes (DGPs) that vary the degree of heterogeneity in compliance and treatment effects, and the correlation between conditional LATEs ($\mathrm{E}\left[Y(1) - Y(0) | D(1) > D(0),\ G = g\right]$) and compliance rates ($\mathrm{E}\left[D(1) - D(0) |\ G = g\right]$).

DGP parameters To simulate a flexible DGP, we use the threshold crossing model representation (Vytlacil, 2002).[39] Let $(\delta_i, \varepsilon_i)' \sim \mathcal{N}(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} \sigma_\delta = 1 & \rho_{\delta\varepsilon} \\ \rho_{\delta\varepsilon} & \sigma_\varepsilon = 1 \end{pmatrix}$$

where δ_i is the latent tendency to receive treatment and ε_i is the baseline untreated potential outcome for individual i. We denote by $\rho_{\delta\varepsilon}$ the correlation coefficient between δ_i and ε_i. The potential treatment indicators are given by:

$$D_i(0) = \mathbb{1}(\Phi_\Sigma(\delta_i) < S_{AT}), \qquad D_i(1) = \mathbb{1}(\Phi_\Sigma(\delta_i) < 1 - S_{NT})$$

where Φ_Σ denotes the cdf of a $\mathcal{N}\left(\vec{0}, \Sigma\right)$, and S_{AT} and S_{NT} represent the share of always-takers and never-takers in the population, respectively. The realized treatment is given by:

$$D_i = D_i(0) \cdot (1 - Z_i) + D_i(1) \cdot Z_i$$

[39] For comparison purposes, we follow Coussens and Spiess (2021) closely in the DGP specifications of their simulations, but deviate in key aspects for reasons that will be explained below.

We also define a covariate X as:

$$X_i = \delta_i + \eta_i$$

where $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$. The covariate X is therefore a noisy predictor of treatment receipt. We also define groups G as the J-quantiles of X:

$$G_i = \mathbb{1}\left(F(X_i) \in \left[\frac{j-1}{J}, \frac{j}{J}\right]\right)$$

So far we have followed the simulation study of Coussens and Spiess (2021), but for the potential outcomes we deviate significantly:

$$Y_i(0) = \varepsilon_i, \qquad Y_i(1) - Y_i(0) = \beta \cdot \left[\alpha \tilde{\pi}_{G(i)} + (1-\alpha)\nu_i\right]$$

where $\tilde{\pi}_G = \pi_G - E_G(\pi_G)$ is the centered compliance rate by group with $G(i)$ representing the group G of individual i and $\nu_i \sim \mathcal{N}(0, \sigma^2_{\pi_G})$. The reason we choose this parametrization of the treatment effect is to generate a significant dependence between compliance rates and treatment effects. Indeed, with this parametrization we have:

$$\begin{aligned}
\sigma^2_{Y(1)-Y(0)} &= \beta^2 \sigma^2_{\pi_G}\left(\alpha^2 + (1-\alpha)^2\right) \\
\operatorname{cov}(\pi_G, Y_i(1) - Y_i(0)) &= \beta \cdot \alpha \cdot \sigma^2_{\pi_G} \\
\operatorname{cor}(\pi_G, Y_i(1) - Y_i(0)) &= \frac{1}{\sqrt{1 + \left(1 - \frac{1}{\alpha}\right)^2}}
\end{aligned}$$

so that β controls the treatment effect heterogeneity and α the dependence between the treatment effect and the compliance rate. Compared to this choice of parametrization, the one chosen in Coussens and Spiess (2021) simulation study generates very little covariance between compliance rates and treatment effects,[40] which is precisely the condition leading to a first-order bias in their estimation strategy.

The Monte Carlo simulations are therefore governed by the following set of parameters:

[40] This comes from the fact that the compliance rate as generated in their DGPs varies non-linearly as a function of δ whereas the LATE depends linearly on δ.

1. N: Sample size
2. J: Number of groups
3. S_{AT}, S_{NT}: Fraction of always-takers and never-takers in the population, respectively
4. $\rho_{\delta\varepsilon}$: correlation between latency to treat and baseline untreated potential. Controls selection into treatment and hence the necessity for instrumentation.
5. σ_η^2: Controls how good of a predictor the groups are for compliance
6. α, β: Control the dependence between treatment effect and compliance as well as the overall treatment effect heterogeneity

Results The selection of DGPs for Monte-Carlo simulations is always a delicate balancing act. We only present 2 classes of DGPs, which we believe showcase some key points we have discussed in the theoretical section. The first DGP (DGP1) illustrates the good properties of our test-and-select estimator in a "best-case scenario" for our estimator, with many groups with 0 compliance alongside groups with large (“strong”) first-stages. Besides demonstrating the potential gains in precision compared to the standard 2SLS estimator, it also highlights the robustness of our estimator to patterns of treatment effect heterogeneity that would bias other alternatives from the literature. The second DGP (DGP2) aims at studying the properties of the various estimators considered in a DGP where no group has 0 compliers, but there are several groups with weak first-stages. This is a setting in which (i) we do not expect significant gains in precision from our estimator and (ii) our selection procedure could lead to some bias depending on the amount of treatment effect heterogeneity. Therefore, this second simulation is another occasion to compare the robustness of our methodology (and its alternatives) to patterns of treatment effect heterogeneity in an adverse DGP.

DGP1: a “best-case scenario” We start by studying a DGP which is an "ideal" application for our method, because 60% of groups have no compliers and the other groups have large compliance rates — in the wording used in section 3, there are only groups with The DGP

parameters are the following:

$$\text{DGP1} \equiv \Big(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\delta\varepsilon} = 0.5, \\ \sigma_\eta = 0.01, \alpha = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\}\Big)$$

It generates the following distribution of compliance rates in the $J = 10$ groups created:

$$\pi_G = (\pi_1 = \pi_2 = \pi_3 = \pi_8 = \pi_9 = \pi_{10} = 0, \pi_4 = \pi_7 \approx 0.25, \pi_5 = \pi_6 \approx 0.99)$$

with an overall compliance rate of 25%. The other important feature of this DGP, encoded by $\alpha = 0.5$, is that there is a significant correlation between compliance and treatment effect. This feature is important because it is the type of treatment effect heterogeneity that can generate bias (with respect to the LATE) in the alternative estimation procedures that have been proposed in the literature (Huntington-Klein, 2020; Coussens and Spiess, 2021; Abadie et al., 2022). In the absence of such correlation, there is no threat of bias neither for our estimator nor these alternatives. Yet as illustrated in section 6, such correlation does exist in real-world applications.

We run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 1.1. We vary treatment effect heterogeneity, and quantify the latter on the x-axis by scaling the standard deviation of $Y(1) - Y(0)$ by the Minimum Detectable Effect (MDE):

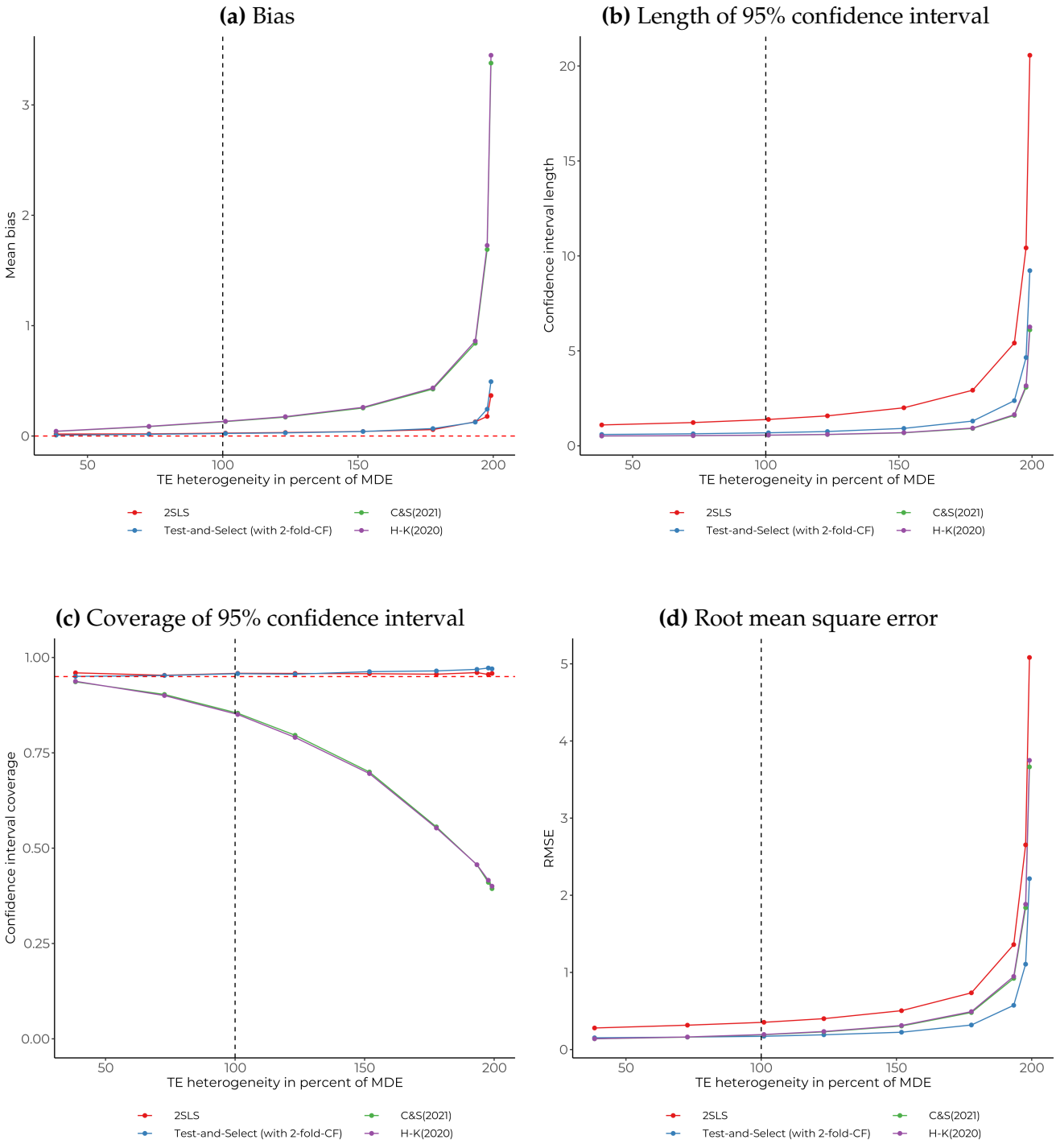
$$x = \frac{\sqrt{V(Y(1) - Y(0))}}{\sqrt{\frac{V(Y(1)) + V(Y(0))}{0.5 \cdot n}} \cdot \frac{q_{0.975} + q_{0.8}}{\pi}}$$

where q_x represents the quantile function of a normal. This re-scaling allows a meaningful quantification of treatment effect heterogeneity, by relating it to a quantity (the MDE) that (i) is a well-known object to most empiricists and (ii) varies with the sample size at a $1/\sqrt{n}$ rate. Recall that at the end of section 3, we highlighted the robustness of our procedure to treatment effect heterogeneity by demonstrating the absence of first-order bias of our estimator when treatment effect heterogeneity is of the order of $1/\sqrt{n}$. The MDE being itself a quantity of this order, quantifying treatment effect heterogeneity with respect to this object allows to get a sense

of whether the level of heterogeneity considered is "small" — i.e., can be deemed of the order $1/\sqrt{n}$ — or "large" and likely to create bias.

Figure 1.1 presents the bias, length and coverage of 95%-CIs, and RMSE of the different estimators considered in these simulations under DGP1. Panel 1.1a highlights the low bias of our estimator up to very large levels of treatment heterogeneity. Estimators based on interacted of weighted instruments display much larger amounts of bias at any level of treatment effect heterogeneity (except zero), as expected. This translates to poor coverage rates for these estimation strategies, when ours covers at the nominal level for any amount of treatment effect heterogeneity — see panel 1.1c. Moreover, panel 1.1b highlights the large decrease in CI length (for all alternative estimation methods) compared to the standard 2SLS. In this DGP, our estimation procedure displays significant gains compared to the standard 2SLS estimator — but as expected less compared to Huntington-Klein (2020) or Coussens and Spiess (2021)'s estimators. Overall, this leads to a domination of our method in terms of RMSE in such a setting — see panel 1.1d. One of the points illustrated in this simulation is that the larger gain in precision associated to Huntington-Klein (2020) or Coussens and Spiess (2021)'s estimators can come along with some bias as long as treatment effect heterogeneity and (conditional) compliance rates are correlated, thus possibly significantly worsening inference on the LATE.

Figure 1.1 – Comparison of estimators with varying treatment effect heterogeneity for DGP1



Notes: This panel shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP1, described in the text. Four different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select estimator with cross-fitting using 2 folds in blue, the re-weighted IV approach suggested by Coussens and Spiess (2021) in green and the interacted IV approach suggest by Huntington-Klein (2020) in purple.

DGP2: introduction of "weak" compliance groups This second DGP is selected for its adverse properties, in order to delineate the robustness frontiers of our method. Indeed, this DGP does not feature any group without compliers. Instead, it introduces several weak compliance groups, which are (as studied in 3) the main source of bias for our estimator. The DGP parameters are the following:

$$\text{DGP2} \equiv \Big(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\delta\varepsilon} = 0.5, \\ \sigma_\eta = 0.5, \alpha = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\}\Big)$$

It generates the following distribution of compliance rates in the $J = 10$ groups created:

$$\pi_G = (\pi_1 = \pi_{10} \approx 0.001, \pi_2 = \pi_9 \approx 0.08, \pi_3 = \pi_8 \approx 0.24, \pi_4 = \pi_7 \approx 0.40, \pi_5 = \pi_6 \approx 0.5)$$

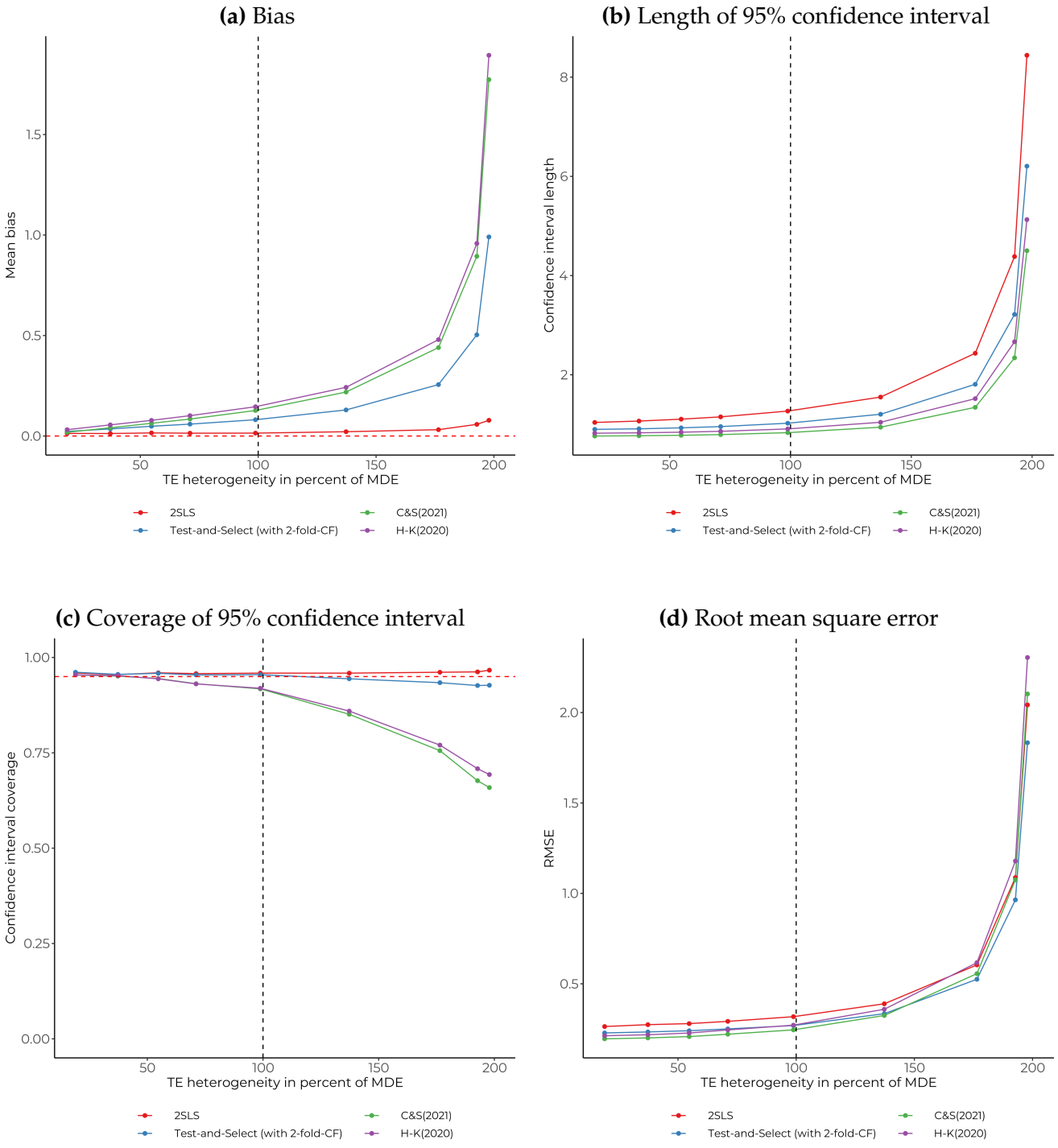
for an overall compliance of 25%, as in DGP1. In the spirit of varying parameters as little as possible across DGPs, we keep the same $\alpha = 0.5$ as in DGP1, which again means that there is a significant correlation between compliance and treatment effects across groups.

We run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 1.2. Compared to what we observed in DGP1 — and as expected from our theoretical results from section 3 — panel 1.2a highlights a much larger bias of our procedure, that grows as treatment effect heterogeneity increases. However, it remains significantly lower compared to the bias of the alternatives proposed in Huntington-Klein (2020) and Coussens and Spiess (2021). Additionally, panel 1.2a conceals the different *distributions* of such estimators compared to ours. Indeed, and as implicitly illustrated in panel 1.2b, Huntington-Klein (2020) and Coussens and Spiess (2021)'s estimators have a lower variance than ours, yielding significantly shorter 95%-CIs. In the absence of any bias, this would unequivocally be synonymous with a better performance. Yet since neither our procedure nor theirs is unbiased, these shorter CIs yield significantly worse coverage properties than our estimator — see panel 1.2c. Indeed, as demonstrated in section 3 our estimator remains unbiased *to the first-order* when treatment effect heterogeneity is moderate — in the sense of being of the same order as the sample variation, or the MDE. Ultimately, such a

property does not guarantee unbiasedness in finite samples — panel 1.2a illustrates this very well — yet it allows for valid inference as long as treatment effect heterogeneity remains moderate. This is precisely what can be seen in panel 1.2c. As treatment effect heterogeneity grows, the coverage of the Test-and-Select estimator's 95%-CIs remains at its nominal level at least up to $x = 100$, and only starts deviating slowly after, while this is not the case of alternative estimators — except for the standard 2SLS of course. Lastly, panel 1.2d shows that the ordering of estimators in terms of RMSE is ambiguous, depending on the level of treatment effect heterogeneity. Yet as we hope to make it clear in this discussion, despite being a standard and useful performance criterion, the RMSE of estimators has to be interpreted with caution here. Indeed, trading off too much bias for gains in precision can very well lead to a decrease in RMSE, yet at the same time be detrimental to the quality of inference by deteriorating the coverage property of standard CIs.[41]

[41] At this point, it is worth mentioning that one could try to correct standard CIs based on estimates of *worst-case* bias of the estimator considered — yielding bias-aware CIs. See Donoho (1996); Armstrong and Kolesár (2018, 2021) for examples of such an approach. This is not explored in this paper, nor in the ones of Huntington-Klein (2020) or Coussens and Spiess (2021). In a companion paper, we study in more details such an alternative.

Figure 1.2 – Comparison of estimators with varying treatment effect heterogeneity for DGP2



Notes: This panel shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP2, described in the text. Four different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select estimator with cross-fitting using 2 folds in blue, the re-weighted IV approach suggested by Coussens and Spiess (2021) in green and the interacted IV approach suggest by Huntington-Klein (2020) in purple.

6 Empirical Applications

6.1 Application to a natural experiment on compulsory schooling laws (Stephens and Yang, 2014)

In this section, we apply our proposed methodology to census data on compulsory schooling laws in the US — as studied in Stephens and Yang (2014).[42] Compulsory schooling laws (that restrict the age at which individuals are allowed to drop out of school) vary across states and time. Assuming that such legal changes occur at random, one can use these variations as an instrument for the amount of schooling of individuals, and therefore identify the causal effect of schooling on wages. In their paper, Stephens and Yang (2014) use an alternative identification strategy, based on parallel trends assumption and a two-way fixed effects model. For the purpose of this application, we propose to make the stronger assumption of random legal changes across states and time.

We start from a sample of 1,175,889 individuals, following the sample selection of Stephens and Yang (2014) except for the fact that the authors choose to focus on white male individuals in their paper while we do not restrict our sample in such a way. Stephens and Yang (2014) justify this restriction by underlining that ethnic minorities and female individuals appear to react less to compulsory schooling laws than male individuals. Motivated by the new estimator proposed in this paper, we suggest making such a selection in a data-driven way, starting from the full sample.

As our main covariate (G in the theory section above), we use an interaction between demographic controls (ethnicity $\times$ sex) $\times$ US census division $\times$ survey year (1960, 1970, 1980). Since we make the assumption that legal changes happen at random, we exclude from our sample the cells defined by G in which there is not any variation in compulsory schooling laws.[43] Indeed, we do not want to identify the effect of compulsory schooling laws on education by

[42] Census data has been used in several other papers to study the effect of education on wages, using compulsory schooling laws as an instrument (Angrist and Krueger, 1991; Acemoglu and Angrist, 2006; Oreopoulos, 2006). We follow Stephens and Yang (2014) for the data cleaning.

[43] This turns out to be necessary once we propose later in this section a natural variation of our estimator that controls non-parametrically for G.

comparing cells in which there has not been any legal changes with some in which there has been some, as such cells are arguably quite different. This restriction is quite stringent, and yields a sample of 171,096 individuals.

Since at this time our methodology only applies to settings with a binary instrument and binary treatment, we need to discretize the original instrument and treatment variables. The original instrument variable in Stephens and Yang (2014) is the number of remaining compulsory years of schooling at age 6 in the state of individuals, at the time they were aged 6. The authors end up discretizing this variable in dummies for whether or not this number is 7, 8 or 9. In order to consider all changes of legislations that imposed to get some high school education, we consider as a single binary instrument the indicator variable that equals one when the number of remaining compulsory years of schooling at age 6 is larger or equal to 7. The original treatment variable is the number of years of schooling completed after age 6. Since some laws require up to 9 years of schooling after age 6, we consider as a treatment variable completing 10 years or more of education. In other words, our treatment variable corresponds to completing some high-school education.

The test-and-select procedure — based on a one-sided t-test on the first stage coefficient within each cell defined by G — tends to select groups of white individuals, as reported in Table 1.2. This confirms the observation of Stephens and Yang (2014) that ethnic minorities tend to react less to the compulsory schooling laws instrument. In fact, these groups often display a negative first-stage, threatening the validity of the identifying assumptions (in particular, the monotonicity assumption) for the LATE.

Table 1.3 reports the results of various estimation procedures applied to the sample described above. Panel (A) reports the results of estimators that do not control in any way for the effect of G on the outcome (log weekly earnings). These include the 2SLS estimator, and the TS estimator with a one-sided t-test with a level of 0.05 or 0.01. We observe that the point estimate of the 2SLS estimator (1.861) differs a bit from the ones of the TS estimators (1.470 or 1.302). Yet the standard errors associated to the TS estimators are smaller — they are reduced by around 12%.

Panel (B) of table 1.3 reports the results of estimators that control (somehow linearly) for

Table 1.2 – Selection probability of G-cells, by demographic group

	Selection proba. ($\alpha = 0.05$)	Selection proba. ($\alpha = 0.01$)
Non-white female	0.39	0.28
Non-white male	0.44	0.33
White female	0.72	0.61
White male	0.67	0.61

Notes: In this application, G is a partition of the population along demographic controls (ethnicity $\times$ sex) $\times$ US census division $\times$ survey year (1960, 1970, 1980). It defines 108 cells, 72 of which are kept in the analysis — those that still contain some variation in our instrument (changes in compulsory schooling laws). This table presents the probability that a cell involving a given demographic group is dropped from the estimation sample once we select based on a one-sided t-test with level 0.05 (first column) or 0.01 (second column).

G. These include in particular the interacted IV estimator (Huntington-Klein, 2020; Coussens and Spiess, 2021) and a version of the estimator proposed in Abadie et al. (2022), the select and interact IV estimator. In fact, these two estimators saturate the first and second stage by including G along with its interactions with Z in the regression. Yet as already mentioned above, the authors show that in such a case, they not identify the LATE, but a convex-weighted average of conditional LATEs. Since G is highly predictive of the outcome in the context of the present application, the variance of these estimators is significantly smaller than the one of the 2SLS estimator and the TS estimator presented in Panel (A). Once we implement the 2SLS and TS estimator after residualizing all variables on G in a first step, the TS estimator display very similar standard errors as the ones of the interacted IV estimators (around 0.150). Moreover, it remains significantly more precise than the 2SLS estimator (0.195).

However, controlling for G linearly as suggested above does not necessarily guarantee that the resulting estimators (2SLS and TS) still target the LATE parameter. In fact, sometimes they could even target a parameter that is a non-convex weighted average of conditional LATEs (Słoczyński, 2022).[44] An alternative is to use another estimator than the 2SLS estimator to

[44]To be precise, under both assumptions of (i) monotonicity and (ii) complete randomization of Z (i.e., Z is independent of the potential outcomes *and* of G), controlling linearly for G does not change the targeted parameter (compared to the situation where we do not control at all for G). Yet in natural or stratified experiments, it could very well be that the distribution of Z varies across along G. In this case (where the second assumption does not hold anymore), the linear control for G yields an estimator of a convex weighted average of conditional LATEs, *yet* with different weights than the natural ones.

control non-parametrically for G. One such estimator has been proposed by Frölich (2007), as an estimator of the LATE when the instrument Z is valid only after conditioning on G. This estimator relies on the following identification results, that states that under unconfoundedness, we can still identify the LATE as the ratio of two weighted average of conditional Intention-To-Treat (ITTs) at the numerator and conditional first-stages at the denominator (see Frölich (2007), theorem 1):

$$\mathrm{E}\left[Y(1) - Y(0) \mid D(1) > D(0)\right] = \frac{\int_{\mathcal{G}} \left(\mathrm{E}[Y \mid G = g, Z = 1] - \mathrm{E}[Y \mid G = g, Z = 0]\right) f_G(g)\mathrm{d}g}{\int_{\mathcal{G}} \left(\mathrm{E}[D \mid G = g, Z = 1] - \mathrm{E}[D \mid G = g, Z = 0]\right) f_G(g)\mathrm{d}g}$$

Since G is discrete in our setting, we can simply use empirical analogs to build a valid estimator of the LATE in our context, that controls non-parametrically for G. We can also construct TS estimators in a similar fashion, by restricting our estimation to the sub-sample of groups selected based on their first-stage. We report the results in table 1.4. One can observe that the resulting point estimates differ quite a lot from the ones previously documented in table 1.3. Indeed, the plain vanilla Frölich estimate is around 0.907 (against 1.86 for the 2SLS estimate without controls). If the instrument were independent from G, then both estimators should target the same LATE parameter. The data does not entirely reject such a scenario since the variance of the Frölich estimator is quite large (1.160). Yet such a difference does suggest that Z might be confounded in the absence of a control for G — which would not be that surprising in this context. If this is the case, then the Frölich estimator is more appropriate. At this stage, this paper does not include a formal discussion of the variance gains of the TS procedure when coupled with the Frölich estimator. Still, in this application, it seems that such a procedure yields considerable variance gains — from 1.16 to 0.604, a reduction by around 48% of the standard errors. The variance of the TS estimator remains larger than the one of the interacted IV estimators in this application. Yet the point estimates of such estimators differ quite a lot from the ones of the Frölich and TS estimators. This suggests the heterogeneity in treatment effect might be such that the interacted IV estimators no longer target the LATE parameter — while the Frölich and TS estimators do. As already discussed in section 3 and 5, this first-order bias of interacted IV estimators can be highly detrimental to the quality of the inference derived based on such estimators. Without the

possibility to de-bias them, and in the absence of a bias-aware procedure for the construction of confidence intervals, it is likely that the said CIs would not cover the LATE parameter at their nominal rate when based on the interacted IV estimators. Indeed, this is due to the failure of such estimators to converge to the LATE at a $\sqrt{n}$ rate (see section 3). On the contrary, the TS procedure yields an asymptotically unbiased estimator, and thus asymptotically valid CIs can be constructed based on this method. Given the significant variance reduction it provides compared to the 2SLS or Frölich alternatives, it appears as the best option to construct tighter, yet asymptotically valid, CIs for the LATE parameter.

Table 1.3 – Comparison of estimation methods

	2SLS	Test and Select (0.05)	Test and Select (0.01)	Interacted IV	Select (0.05) & Interacted IV
		A. Without controlling for G (demographic controls)			
D (educ. $\geq$ some high-school)	1.861	1.470	1.302		
	(0.365)	(0.320)	(0.312)		
	[1.145, 2.578]	[0.843, 2.098]	[0.691, 1.913]		
First-stage coef.	0.523	0.518	0.513		
%. sample drop.	0	28.6	32.7		
N	171 096	122 150	115 159		
		B. Controlling (linearly) for G (demographic controls)			
D (educ. $\geq$ some high-school)	1.370	1.143	1.130	1.348	1.149
	(0.195)	(0.150)	(0.154)	(0.182)	(0.149)
	[0.989, 1.751]	[0.849, 1.437]	[0.828, 1.432]	[0.991, 1.705]	[0.858, 1.441]
First-stage coef.	0.053	0.083	0.087	0.088	0.103
%. sample drop.	0	28.6	32.7	0	28.6
N	171 096	122 150	115 159	171 096	122 150

Notes: Standard errors (clustered at the demographic control (ethnicity $\times$ sex) $\times$ birth state $\times$ year of birth level) in parenthesis, 95% confidence intervals in brackets. We report estimates of the effect of having some high-school education (or more) on log weekly earnings.

Table 1.4 – Comparison of estimation methods (continued)

	Frölich (2007)	TS (0.05) & Frölich (2007)	TS (0.01) & Frölich (2007)	Interacted IV	Select (0.05) & Interacted IV
	C. Controlling (non-parametrically) for G (demographic controls)				
D (educ. $\geq$ some high-school)	0.907	0.791	0.776	1.348	1.149
	(1.16)	(0.604)	(0.576)	(0.182)	(0.149)
	[−1.367, 3.181]	[−0.392, 1.975]	[−0.353, 1.905]	[0.991, 1.705]	[0.858, 1.441]
First-stage coef.	0.064	0.094	0.097	0.088	0.103
%. sample drop.	0	28.6	32.7	0	28.6
N	171 096	122 150	115 159	171 096	122 150

Notes: Standard errors (clustered at the demographic control (ethnicity $\times$ sex) $\times$ birth state $\times$ year of birth level) in parenthesis. We report estimates of the effect of having some high-school education (or more) on log weekly earnings.

6.2 Application to a large-scale controlled experiment on job search counseling (Behaghel et al., 2014)

In this section, we apply our proposed methodology to a large-scale labor market experiment on job search counseling studied in Behaghel et al. (2014). This randomized controlled trial aimed at measuring (and comparing) the impact of intensive job search counseling delivered either by public (CVE) or private (OPP) providers. Among a pool of job seekers at risk of long-term unemployment, the ones assigned to either of those two treatment arms were eligible to receive counseling from advisors whose caseload was reduced (on average) from 120 to 40 job seekers.

We restrict our sample to control individuals and job seekers assigned to the intensive counseling program of public providers (CVE) — in order to focus on a single treatment arm, while keeping the largest number of observations possible. We are left with a sample of 113,738 job seekers for our analysis.[45] We focus on measuring the effect on the number of days spent as unemployed during the year after the date of the assignment.

We have access to a rich set of individual covariates from the administrative data of the French Public Employment Services (PES) — e.g. the age, region, marital status, number of children, nationality, area of residence, occupation he/she is looking for, qualification, level of education, reasons for unemployment registration (fired, quit, economical downsizing) etc. In order to fit into our framework, we build a synthetic variable that aims at summarizing the predictive power of those covariates for compliance behavior. In order to do so, we grow a random forest on the subsample of assigned individuals, whose objective is to best predict the treatment variable (entering into the CVE program) based on observables. In future research, we would like to investigate the best way to use high-dimensional covariates in our setting, taking into account the use of such prediction algorithm in our analysis. At this stage, in order to best approximate the existence of an exogenous partition of the population — as assumed in our theoretical framework — we randomly split our sample in two halves, and build two distinct

[45] Compared to the original published paper of Behaghel et al. (2014), our sample is larger than the one used in their main analysis. This is because they had to restrict their analysis to job seekers that were eligible to both programs (CVE and OPP). There was a higher number of job seekers in the experiment that were eligible to CVE (and not necessarily to OPP), hence our bigger sample.

prediction models. The prediction function estimated in split 1 is then applied to split 2, and vice versa. From there, within each split, we create 500 quantiles of this compliance score, that are going to be used as the main covariates in our analysis. This data-splitting and cross-fitting allows us to consider this covariate as essentially exogenous, as the prediction function used in each split to create those quantiles does not depend on the realizations of the data within the split.

Table 1.5 presents the heterogeneity of the (conditional) LATE along quartiles of the predicted compliance rate. We highlight two main points from this table. First, it provides evidence that we successfully captured some heterogeneity in compliance behavior, as the first-stage coefficients estimated in each quartile of predicted compliance are highly heterogeneous, from an average take-up rate of 18.4% in the first quartile to a rate of 49.5% in the last quartile. This demonstrates the ability of prediction models to capture heterogeneous compliance behaviors along observables in real-world datasets. The second fact worth noticing is the covariance between the conditional LATEs estimated in each quartile, and the conditional take-up rates mentioned above. Indeed, LATE estimates vary considerably from the first to the last quartile of compliance, going from around +17 to -16 days of unemployment. This can easily be rationalized by a Roy model in which job seekers self-select into treatment (when assigned to) based on their expected gains from such program. However, it is critical to document such a pattern in a real-world dataset. Indeed, we highlighted in our theoretical derivations and Monte-Carlo simulations that our proposed Test-and-Select estimator was more robust to such covariance between compliance rates and treatment effects (compared to alternative estimators). Yet this robustness could be deemed vain if real datasets failed to present significant covariance between treatment effects and compliance rates.

Table 1.6 then compares the different estimation methods for the LATE parameter. Its first column presents the standard 2SLS estimate, at around -5.3 days of unemployment. The second and third columns of the table present the results obtained when applying our methodology when testing either at the 0.05 or 0.01 level in the selection step. Mechanically, using a 0.01 level leads to a larger fraction of the sample being dropped. Both alternatives do not yield

Table 1.5 – Heterogeneity across quartiles of predicted compliance

	Q1	Q2	Q3	Q4
Constant	186.361	216.297	233.778	264.426
	(1.146)	(1.128)	(1.103)	(1.039)
	[184.114, 188.607]	[214.087, 218.507]	[231.616, 235.940]	[262.390, 266.462]
Treatment (CVE)	16.916	−0.304	−10.758	−15.993
	(8.652)	(5.872)	(4.210)	(2.785)
	[−0.042, 33.875]	[−11.814, 11.205]	[−19.010, −2.507]	[−21.453, −10.534]
First-stage coef.	0.184	0.261	0.356	0.495
	(0.004)	(0.004)	(0.004)	(0.005)
N	28 272	28 489	28 375	28 602

Notes: Robust standard errors in parenthesis, 95% confidence intervals in brackets. The dependent variable is the number of days spent as unemployed during the year after date of the assignment. Each column reports the 2SLS estimates from a subsample restricted to observations among a given quartile of predicted compliance score.

significant gains in variance, which can be explained by the very moderate increase in the average take-up rate. The last two columns present estimates based on alternative methodologies. The second to last column corresponds to an interacted instrument estimation approach as suggested in Huntington-Klein (2020), while the last column presents an estimate based on the weighted-instrument strategy suggested in Coussens and Spiess (2021), with our compliance score estimated by random forest as the weight. Both are (as expected) very similar, with significant gains in variance along with relatively larger deviation of their point estimates from the 2SLS one (compared to our methodology). This is not surprising as those methodologies target a weighted average of conditional LATEs where populations with higher compliance rates get a larger weight. Therefore, as Table 1.5 documented the covariance between compliance rates and larger (negative) effects on days spent in unemployment, we would expect those estimators to be centered on larger estimands, which is what the results in Table 1.6 seem to confirm. Notice that the version of our TS procedure at the 0.01 level does show a similar pattern, yet to a lesser extent, as expected from our theoretical results.

Table 1.6 – Comparison of estimation methods

	2SLS	Test and Select (0.05)	Test and Select (0.01)	H.-K. (2020)	C. & S. (2021)
Constant	224.993	226.057	229.268	177.338	162.682
	(0.569)	(0.576)	(0.602)	(0.781)	(0.930)
	[223.878, 226.109]	[224.928, 227.186]	[228.089, 230.447]	[175.807, 178.868]	[160.859, 164.505]
Treatment (CVE)	−5.294	−5.664	−8.796	−10.600	−11.134
	(2.352)	(2.356)	(2.379)	(2.092)	(2.097)
	[−9.903, −0.684]	[−10.281, −1.046]	[−13.459, −4.132]	[−14.701, −6.500]	[−15.243, −7.024]
First-stage coef.	0.326	0.329	0.337		
%. sample drop.	0	2.4	9.8	0	0
N	113 738	110 998	102 560	113 738	113 738

Notes: Robust standard errors in parenthesis, 95% confidence intervals in brackets. The dependent variable is the number of days spent as unemployed during the year after the date of the assignment. The first model reports the results of 2SLS estimation of the full sample. The second and third columns report the results of estimating the LATE on a subsample selected based on a first testing step (implemented using data splitting and cross-fitting as described in the main text). The second column corresponds to a selection rule based on a 0.95 level t-test, and the third column corresponds to a selection rule based on a 0.99 level t-test. The penultimate column corresponds to an interacted instrument estimation approach as suggested in Huntington-Klein (2020). The last column presents an estimate based on the weighted-instrument strategy suggested in Coussens and Spiess (2021), with our compliance score estimated by random forest as the weight.

7 Conclusion

In this paper, we consider a simple and intuitive way to exploit heterogeneity in compliance rates along observable characteristics in order to improve the estimation of the LATE in experiments with imperfect compliance. We start by underlining the fact that excluding non-compliant sub-populations from the analysis does not affect the estimand identified while allowing to reduce considerably the variance of a hypothetical oracle estimator that would exclude observations from such population without any exclusion mistakes. Quite naturally, this result on precision gains extends asymptotically to a feasible estimator that would identify such non-compliant groups by t-testing the first-stage coefficient as long as we consider standard asymptotic sequences in which compliance rates per group are either zero or fixed with n and well-separated from 0. Yet such asymptotic results are likely to yield unsatisfactory approximations of our estimator's behavior in finite samples. Therefore, we next consider weak-IV-like asymptotic sequences in which some groups display local-to-zero compliance rates — i.e., their first-stage coefficient decreases at the $1/\sqrt{n}$ rate, making them difficult to distinguish from non-compliant groups in samples of any size. We provide sufficient conditions — in particular, restrictions on treatment effect heterogeneity — for our estimator to remain first-order unbiased for the LATE under such asymptotic sequences. We discuss the interpretability of such conditions in applied work and compare the performance of our estimator to alternative procedures recently proposed in the literature, which exploit first-stage heterogeneity differently from us. The main takeaway from this discussion is that our estimator appears more robust to treatment effect heterogeneity, mainly because it exploits specific patterns of compliance rates heterogeneity — namely, the presence of non-compliant groups. The cost of such robustness is limited gains in precision when the non-compliant sub-population cannot be described accurately by observable characteristics. In light of our theoretical findings, we explore the finite sample performance of our estimator in Monte-Carlo simulations and in an application on a large RCT on job search counseling. Both our simulations and the application confirm the higher robustness of our estimator to treatment effect heterogeneity. The potential for precision gains is also clearly highlighted in Monte-Carlo simulations.

The econometrics literature on the use of first-stage heterogeneity in LATE estimation is very recent and thus still quite active and promising. As an example, in a follow-up project (joint with X. D'Hautefœuille) we reflect on the setting studied in this paper under the milder restriction of *bounded* treatment effect heterogeneity. We consider the use of bias-aware inference techniques, that have received a renewed attention in the recent econometric literature on treatment effect estimation. In the case of LATE estimation with heterogeneous first-stages across groups defined by covariates, this assumption of bounded treatment effect heterogeneity yields a set of restrictions on the relationship between the Intention-to-Treat (ITT) and the first-stage statistics within each group — which can then be used to construct bias-aware CIs on the LATE, with the hope that such procedure could yield a more precise inference than standard approaches. Aside from the benefits from taking first-stage heterogeneity into account in estimation and inferential procedures — as in this paper and its follow-up — we believe that these insights could be used for the design of experiments with imperfect compliance. We plan on investigating this in future research.

A Appendix

A.1 Proofs of main results

Proof. Lemma 1.

Let G be a binary covariate partitioning the population such that:

- the share of compliers in groups $G = 0$ and $G = 1$ are respectively given by $\pi^0 = 0$ and $\pi^1 > 0$. We denote by $\hat{\pi}^0$ and $\hat{\pi}^1$ the first-stage estimators in each of those two groups.
- the LATE in group $G = 1$ is denoted $LATE_1$. Note that is matches the LATE in the overall population since there are not any compliers in group $G = 0$.
- in group $G = 0$, we have:

$$B_{AT-NT} \equiv E\left[Y(1)|G = 0, D(1) = D(0) = 1\right] - E\left[Y(0)|G = 0, D(1) = D(0) = 0\right] \neq 0$$

The last point states that the average outcome of always-takers — characterized by $D(1) = D(0) = 1$, and for whom we always observe $Y(1)$ — is different from the average outcome of never-takers — characterized by $D(1) = D(0) = 0$, and for whom we always observe $Y(0)$.

First of all, notice that group $G = 1$ is selected with probability tending to 1 as n goes to infinity (by consistency of the t-test against alternatives well separated from 0). With probability tending to $(1 - \alpha)$ — where α is the level of the t-test used for selection — group $G = 0$ is not selected. See lemma 2 for a proof of these statements. Therefore, the event (resulting from our unilateral t-test on group first-stages) we are interested in is:

$$\{\text{Group G=0 is selected}\} \Leftrightarrow \left\{\mathbb{1}\left\{\sqrt{n^0} \cdot \frac{\hat{\pi}^0}{\hat{\sigma}^{\hat{\pi}^0}} \; > \; q_{1-\alpha}\right\} = 1\right\}$$

With probability tending to $(1 - \alpha)$, only group $G = 1$ is selected. The event determining whether group 1 is selected alone or not does not depend on the observations in this group. Therefore, the 2SLS estimator computed on observations of group $G = 1$ alone has an asymptotic distribution *conditonal* on the event {Group G=0 is selected} that remains the same as its unconditional asymptotic distribution. By standard results on 2SLS estimation we get that the

standard 2SLS estimator computed on observations from subgroup $G = 1$ (denoted $\widehat{LATE}_1$) will be asymptotically normal and centered on $LATE_1$:

$$\sqrt{n^1} \cdot \left(\widehat{LATE}_1 - LATE_1 \right) \xrightarrow[n\to\infty]{} \mathcal{N}(0, V^1)$$

Yet when both group $G = 0$ and $G = 1$ are selected — with asymptotic probability α — the 2SLS estimator computed on both groups (denoted $\widehat{LATE}$) satisfies:

$$\begin{aligned}
&\sqrt{n} \cdot \left(\widehat{LATE} - LATE \right) \\
&= \sqrt{n} \cdot \left(\widehat{LATE} - LATE_1 \right) \\
&= \sqrt{n} \cdot \left(\frac{\widehat{ITT_1}}{\hat{\pi}^1} - LATE_1 - \underbrace{\frac{\widehat{ITT_1}}{\hat{\pi}^1} \cdot \frac{\hat{P}_0 \cdot \hat{\pi}^0}{\hat{P}_0 \cdot \hat{\pi}^0 + \hat{P}_1 \cdot \hat{\pi}^1}}_{\equiv A} + \underbrace{\frac{\hat{P}_0 \cdot \hat{\pi}^0 \cdot \widehat{ITT_0}}{\hat{P}_0 \cdot \hat{\pi}^0 + \hat{P}_1 \cdot \hat{\pi}^1}}_{\equiv B} \right) \\
&= \sqrt{n} \cdot \left(\frac{\widehat{ITT_1}}{\hat{\pi}^1} - LATE_1 \right) - \sqrt{n} \cdot A + \sqrt{n} \cdot B
\end{aligned}$$

where $\hat{P}_g \equiv \hat{\mathrm{P}}[G = g] = n_g/n$ and $\widehat{ITT_g}$ denotes the difference-in-means estimator of the intention-to-treat estimand ($E[Y|Z = 1] - E[Y|Z = 0]$) in group $G = g$. If we were reasoning unconditionally — i.e., without conditioning on the event {Group G=0 is selected} — then we would have that both A and B have distributions centered on 0 — by Slutsky and the continuous mapping theorem, since $\sqrt{n} \cdot \hat{\pi}^0 \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{\hat{\pi}^0})$. Thus $\widehat{LATE}$ would be $\sqrt{n}$-consistent for the LATE. Yet, we are interested in the distribution of $\widehat{LATE}$ conditional on the event {Group G=0 is selected}, which is equivalent to conditioning on $\sqrt{n} \cdot \hat{\pi}^0$ being larger than a given threshold t. We trivially have:

$$\sqrt{n} \cdot \hat{\pi}^0 \mid \sqrt{n} \cdot \hat{\pi}^0 > t \xrightarrow[n\to\infty]{d} \mathcal{N}(0, LB = t, V_{\hat{\pi}^0})$$

where $\mathcal{N}(0, LB = t, V_{\hat{\pi}^0})$ denote the distribution of a truncated normal distribution $\mathcal{N}(0, V_{\hat{\pi}^0})$ with lower bound t. Such distribution is not centered on 0. Therefore, since $\widehat{ITT_1}$ does not go to 0, we already have that our first bias term A does not vanish anymore. This is a first source of

first-order bias in the estimation of the LATE with this naïve pre-testing procedure. This one is quite intuitive: as our pre-test tends to select cases in which we overestimate the share of compliers in group $G = 0$, we tend to overestimate the overall share of compliers, and thus this shrinks the estimator towards 0.

However, there is potentially a second source of bias that comes from the non causal comparison between always-takers and never-takers in group $G = 0$. Indeed, since there are not any compliers in this group, having a large first-stage in $G = 0$ necessarily means that there is an imbalance between the share of always takers and the share of never-takers in this sub-sample. If we do not condition on the size of the estimated first-stage coefficient $\hat{\pi}^0$, then we still have that those shares are balanced on average, and thus we have $\widehat{ITT^0} \xrightarrow[n\to\infty]{p} 0$ and, by Slutsky's lemma $\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \tilde{V}_0)$. Yet once we condition on the estimated first-stage coefficient, the probability limit of $\widehat{ITT^0}$ and the limiting distribution of $\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0}$ are quite different. Indeed, we have:

$$\widehat{ITT^0} \mid \hat{\pi}^0 = f \xrightarrow[n\to\infty]{p} f \cdot B_{AT-NT}$$

Hence once we turn to the study of the limiting distribution of $\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0}$, we get:

$$\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0} \mid \sqrt{n} \cdot \hat{\pi}^0 > t \xrightarrow[n\to\infty]{d} \mathcal{N}(0, LB = t, V_{\hat{\pi}^0}) \cdot B_{AT-NT}$$

If $B_{AT-NT} = 0$, then this limiting distribution becomes degenerate at 0, and the second bias term B is null. Yet if $B_{AT-NT} = 0$, then this additional term B is not centered at 0, and therefore it adds an additional first-order bias to the estimator $\widehat{LATE}$. Once again, this is intuitive as this second bias term B comes from the fact that in group $G = 0$, we end up comparing always-takers with never-takers once we condition on the estimated first-stage $\hat{\pi}^0$ to be larger than a threshold. This is not an issue when the expected outcome of always takers and never-takers is the same ($B_{AT-NT} = 0$), as this difference will concentrate around zero in this case. This is not the case if the expected outcome of always-takers and never-takers differ ($B_{AT-NT} \neq 0$), in which case their comparison leads to the introduction of a first-order bias.

□

Proof. Proposition 1.

Proposition 1.1 We'll closely follow the proof of lemma 7, that presents the asymptotic distribution of the usual 2SLS/Wald estimator. The steps are essentially identical, but for an additional conditioning on S_{G_i}, the selection dummy indicating whether the covariate-based group individual i belongs to (denoted by G_i) has been selected. This is indicated in vector $S \in \{0,1\}^{|\mathcal{G}|}$. S_{G_i} is merely the G_i^{th} line of the vector S. Let us also us the following notation:

- $\mathcal{G}_S = \{\text{all groups with strong fist stage}\}$
- $\mathcal{G}_0 = \{\text{all groups with zero fist stage}\}$

We do not consider groups with weak first-stages at this point, as proposition 1 focuses on standard asymptotics in order to illustrate the potential for gains in precision from selection. Notice that Propositions 1.1 and 1.2 are developed under a conditioning on the value of the selection vector S. This is key to our reasoning, as this conditioning allows us to study separately the randomness of the estimation sample, and the one coming from the selection step.
Consider a given (fixed, deterministic) selection process $S \in \mathcal{S}_{\text{strong}}$. We know that asymptotically, it cannot be that a group with a strong first-stage is not selected. Hence there are only two main cases we need to consider:

1. $\{\forall g \in \mathcal{G}_S, S_g = 1\} \cap \{\forall g \in \mathcal{G}_0, S_g = 0\}$
2. $\{\forall g \in \mathcal{G}_S, S_g = 1\} \cap \{\exists g \in \mathcal{G}_0, S_g = 1\}$

The various components of $\hat{\tau}(S)$ are:

$$\begin{aligned}
\hat{A} &= \left(\sum_i Z_i S_{G_i}\right)^{-1} \sum_i Z_i S_{G_i} Y_i, && A = E[Y|Z=1, S_G=1] \\
\hat{B} &= \left(\sum_i ((1-Z_i) S_{G_i})\right)^{-1} \sum_i (1-Z_i)\, S_{G_i} Y_i, && B = E[Y|Z=0, S_G=1] \\
\hat{C} &= \left(\sum_i Z_i S_{G_i}\right)^{-1} \sum_i Z_i S_{G_i} D_i, && C = E[D|Z=1, S_G=1] \\
\hat{D} &= \left(\sum_i ((1-Z_i) S_{G_i})\right)^{-1} \sum_i (1-Z_i)\, S_{G_i} D_i, && D = E[D|Z=0, S_G=1] \\
&\Rightarrow LATE = \frac{A-B}{C-D} \\
&\Rightarrow \hat{\tau}(S) = \frac{\hat{A}-\hat{B}}{\hat{C}-\hat{D}}
\end{aligned}$$

Notice that the fact that $LATE = \frac{A-B}{C-D}$ comes from the fact that no matter the selection procedure $S \in \mathcal{S}_{\text{strong}}$ considered, the only groups that might be excluded are groups without any compliers. Therefore we get:

$$\begin{aligned}
\text{LATE} = \;& \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), S_G=1] \cdot \overbrace{\mathrm{P}[S_G=1|D(1)>D(0)]}^{=1} \\
&+ \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), S_G=0] \cdot \underbrace{\mathrm{P}[S_G=0|D(1)>D(0)]}_{=0} \\
&= \mathrm{E}[Y(1)-Y(0)|D(1)>D(0), S_G=1] \\
&= \frac{E[Y|Z=1,S_G=1]-E[Y|Z=0,S_G=1]}{E[D|Z=1,S_G=1]-E[D|Z=0,S_G=1]} \quad \text{(by standard identification result for LATE)}
\end{aligned}$$

In exactly the same way as the proof of lemma 7, we have:

$$
\begin{aligned}
a_i &= \frac{Z_i S_{G_i}(Y_i - E[Y|Z=1, S_G=1])}{E[ZS]} \\
b_i &= \frac{(1-Z_i) S_{G_i}(Y_i - E[Y|Z=0, S_G=1])}{E[(1-Z)S_G]} \\
c_i &= \frac{Z_i S_{G_i}(D_i - E[D|Z=1, S_G=1])}{E[ZS_G]} \\
d_i &= \frac{(1-Z_i) S_{G_i}(Y_i - E[D|Z=0, S_G=1])}{E[(1-Z)S_G]}
\end{aligned}
$$

Therefore we get:

$$
\begin{aligned}
\psi_{\hat{\tau}(S),i} &= \frac{(a_i - b_i) - LATE \cdot (c_i - d_i)}{C_i - D_i} \\
&= \frac{1}{p_{C,S_G=1}} \left(\frac{Z_i S_{G_i}(Y_i - E[Y|Z=1, S_G=1])}{E[ZS_G]} - \frac{(1-Z_i) S_{G_i}(Y_i - E[Y|Z=0, S_G=1])}{E[(1-Z)S_G]} \right. \\
&\qquad \left. - LATE \cdot \left(\frac{Z_i S_{G_i}(D_i - E[D|Z=1, S_G=1])}{E[ZS_G]} - \frac{(1-Z_i) S_{G_i}(D_i - E[D|Z=0, S_G=1])}{E[(1-Z)S_G]} \right) \right) \\
&= \frac{1}{p_{C,S_G=1}} \left(\frac{1}{E[ZS_G]} Z_i S_{G_i} \cdot (\varepsilon_i - E[\varepsilon|Z=1, S_G=1]) - \frac{1}{E[(1-Z)S_G]} (1-Z_i) S_{G_i} \cdot (\varepsilon_i - E[\varepsilon|Z=0, S_G=1]) \right)
\end{aligned}
$$

where $\varepsilon \equiv Y - LATE \cdot D$ is the structural error term of the second stage, and $p_{C,S_G=1} = \mathrm{E}[D(1) - D(0)|S_G = 1]$ is the share of compliers among the selected. As expected from an influence function, one can check that $E[\psi_{\hat{\tau}(S),i}] = 0$. It follows that asymptotically,

$$
\sqrt{n_{(E)}}(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(0, V^{\hat{\tau}(S)})
$$

where $V^{\hat{\tau}(S)} = V(\psi_{\hat{\tau}(S),i})$ equals:

$$
\begin{aligned}
V(\psi_{\hat{\tau}(S),i}) &= E[\psi^2_{\hat{\tau}(S),i}] \\
&= \frac{1}{p^2_{C,S_G=1}} \left(\frac{1}{E[ZS_G]} E[(\varepsilon - E[\varepsilon|Z=1, S_G=1])^2 | Z=1, S_G=1] \right. \\
&\qquad \left. + \frac{1}{E[(1-Z)S_G]} E[(\varepsilon - E[\varepsilon|Z=0, S_G=1])^2 | Z=0, S_G=1] \right)
\end{aligned}
$$

We also have $Z \perp S_G$ (because $Z \perp G$ and S is deterministic as we condition on it), so that:

$$\begin{aligned}
E[ZS_G] &= p \cdot p_{S_G} \\
E[(1-Z)S_G] &= (1-p) \cdot p_{S_G} \\
\pi &= p_{C,S_G=1} \cdot p_{S_G} + p_{C,S_G=0} \cdot (1 - p_{S_G}) = p_{C,S_G=1} \cdot p_{S_G} \\
\Rightarrow V(\psi_{\hat{\tau}(S),i}) &= \frac{p_{S_G}}{\pi^2} \left(\frac{1}{p} E[(\varepsilon - E[\varepsilon|Z=1,S_G=1])^2 | Z=1, S_G=1] \right. \\
&\qquad \left. + \frac{1}{1-p} E[(\varepsilon - E[\varepsilon|Z=0,S_G=1])^2 | Z=0, S_G=1] \right)
\end{aligned}$$

where $p_{S_G} \equiv \Pr[S_G = 1]$.

Proposition 1.2 From lemma 7, and from proposition 1.1 we have that:

$$\begin{aligned}
V^{TSLS} &= \frac{1}{\pi^2} \left(\frac{1}{p} V[\varepsilon|Z=1] + \frac{1}{1-p} V[\varepsilon|Z=0] \right) \\
V^{\hat{\tau}(S)} &= \frac{1}{\pi^2} \left(\frac{p_{S_G}}{p} V[(\varepsilon|Z=1,S_G=1] + \frac{p_{S_G}}{1-p} V[(\varepsilon|Z=0,S_G=1] \right)
\end{aligned}$$

Therefore, we only need to prove that:

$$V[\varepsilon|Z=z] \geq p_{S_G} \cdot V[\varepsilon|Z=z,S_G=1]$$

This is proven below:

$$\begin{aligned}
V(\varepsilon|Z=z) &= E\left[V(\varepsilon|Z=z,S_G)|Z=z\right] + V\left(E\left[\varepsilon|Z=z,S_G\right]|Z=z\right) \\
&\geq E\left[V(\varepsilon|Z=z,S_G)|Z=z\right] \\
&\geq p_{S_G} \cdot V(\varepsilon|Z=z,S_G=1)
\end{aligned}$$

where the first equality follows from the law of total variance, and first and second inequalities follow from the fact that variances are always positive or null (in degenerate cases).
Therefore, V^{TSLS} has been shown to be a linear combination (with positive coefficients) of terms greater or equal than the ones appearing in $V^{\hat{\tau}(S)}$, proving the proposition 1.2.

Proposition 1.3 In order to properly study the asymptotic distribution of $\hat{\tau}_T = \hat{\tau}(\hat{S}_{(T)})$, we need to take a step back and study the distribution of $\hat{S}_{(T)}$, the vector of selection indicators estimated in the test sample $\mathcal{I}_T$. We can focus on any single indicator $\hat{S}_{g,(T)}$, the g^{th} line of vector $\hat{S}_{(T)}$, which is defined as follows:

$$\hat{S}_{g,(T)} \equiv \mathbb{1}\left\{\hat{\pi}^g_{(T)} > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(T)}}} \cdot q_{1-\alpha}\right\}$$

where $n^g_{(T)}$ is the number of observations in group g in sample $\mathcal{I}_T$, $\hat{\pi}^g_{(T)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance of $\hat{\pi}^g_{(T)}$. Notice that $\hat{\pi}^g_{(T)}$ is asymptotically linear, as following lemma 6 we have:

$$\begin{aligned}
&\sqrt{n^g_{(T)}} \cdot [\hat{\pi}^g - \pi^g] \\
&= \sqrt{n^g_{(T)}} \cdot \left[\frac{\sum_i Z_i D_i}{\sum_i Z_i} - \frac{\sum_i (1-Z_i) D_i}{\sum_i (1-Z_i)} - (\mathbb{E}[D \mid Z=1] - E[D \mid Z=0])\right] \\
&= \frac{1}{\sqrt{n^g_{(T)}}} \cdot \left[\sum_{i=1}^{n^g_{(T)}} \underbrace{\left(\frac{Z_i\,(D_i - \mathbb{E}[D \mid Z=1]}{E[Z]} + \frac{(1-Z_i)\cdot(D_i - \mathbb{E}[D \mid Z=0])}{1-\mathbb{E}[Z]}\right)}_{\equiv \tilde{\psi}^g_i}\right] \\
&= \frac{1}{\sqrt{n^g_{(T)}}} \cdot \sum_{i=1}^{n^g_{(T)}} \tilde{\psi}^g_i
\end{aligned}$$

Our estimator $\hat{\tau}_T$ depends on the selection variables stacked in $\hat{S}_{(T)}$. Indeed, we have:

$$\sqrt{n_{(E)}}(\hat{\tau}_T - LATE) = \frac{1}{\sqrt{n_{(E)}}} \sum_i \psi_{\hat{\tau}_T, i}$$

where the expression of the influence function is given by:

$$\begin{aligned}
\psi_{\hat{\tau}_T,i} =& \frac{1}{p_{C,\hat{S}_{G,T}=1}} \left(\frac{1}{E[Z\hat{S}_{G,T}]} Z_i \hat{S}_{G_i,T} \cdot (\varepsilon_i - E[\varepsilon | Z=1, \hat{S}_{G,T}=1]) \right. \\
&\left. - \frac{1}{E[(1-Z)\hat{S}_{G,T}]} (1-Z_i)\hat{S}_{G_i,T} \cdot (\varepsilon_i - E[\varepsilon | Z=0, \hat{S}_{G,T}=1])\right)
\end{aligned}$$

The above display makes it clear that the $\psi_{\hat{\tau}_T,i}$'s of individuals within a given group g are dependent, as they all depend on $\hat{S}_{g,T}$, the selection indicator computed in the test sample $\mathcal{I}_T$. Yet the fact that this variable is computed in a different sample allows us to disentangle the randomness of $\hat{\tau}_T$ *conditional* on the selection vector $\hat{S}_T$, and the randomness of the selection process $\hat{S}_T$ itself. Conditioning on the selection vector $\hat{S}_T$ re-establishes independence across the $\psi_{\hat{\tau}_T,i}$'s, and we are back to the case studied in proposition 1.1 and 1.2. Now let us define:

$$\hat{T}_E \equiv \sqrt{n_{(E)}} \cdot \frac{\hat{\tau}_T - LATE}{\sqrt{\hat{V}(\tau(\hat{S}_T))}}$$

where $V^{\hat{\tau}_E}$ is the asymptotic variance of $\hat{\tau}_E = \hat{\tau}(\hat{S}_T)$. Now, turning to the study of the characteristic function of $\hat{T}_E$ conditional on $\hat{S}_T$, we have:

$$\begin{aligned}
\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T] &= \sum_{S\in\{0,1\}^{|\mathcal{G}|}} \mathbb{1}\{\hat{S}=S\}\cdot \mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T=S] \\
&\xrightarrow[n\to\infty]{p} \sum_{S\in\mathcal{S}_{strong}} \mathbb{1}\{\hat{S}=S\}\cdot e^{-t^2/2} + \sum_{S\notin\mathcal{S}_{strong}} 0\cdot \mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T=S]
\end{aligned}$$

Indeed, by proposition 1.1 we have that for $\hat{S}_T \in \mathcal{S}_{strong}$, $\hat{T}_E$ converges to a $\mathcal{N}(0,1)$. And by consistency of the t-test against any alternative well separated from 0, we have that all groups with strong first-stages are selected asymptotically, implying: $\forall S \notin \mathcal{S}_{strong},\ \mathbb{1}\{\hat{S}_T = S\} \xrightarrow[n\to\infty]{p} 0$, hence the second line of the above display (by continuous mapping theorem).
Notice that by Jensen inequality: $|\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T]| \leq \mathrm{E}[|e^{it\hat{T}_E}||\hat{S}_T] = 1$, hence by the dominated convergence theorem we get:

$$\begin{aligned}
\mathrm{E}[e^{it\hat{T}_E}] = \mathrm{E}\left[\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T]\right] &\xrightarrow[n\to\infty]{p} E\left[\sum_{S\in\mathcal{S}_{strong}} \mathbb{1}\{\hat{S}=S\}\cdot e^{-t^2/2} + \sum_{S\notin\mathcal{S}_{strong}} 0\right] \\
&= e^{-t^2/2} \quad \text{(characteristic function of a } \mathcal{N}(0,1)\text{)}
\end{aligned}$$

By Jensen inequality we have: $|\mathrm{E}[e^{it\hat{T}_E}]| \leq \mathrm{E}[|e^{it\hat{T}_E}|] = 1$ and since convergence in probability and boundedness (in $\mathbb{C}$) imply convergence in $\mathcal{L}^1$, we have:

$$\mathrm{E}\left[\left|\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T] - e^{-t^2/2}\right|\right] \xrightarrow[n\to\infty]{} 0$$

By Jensen inequality again, we have:

$$\begin{aligned}\left|\mathrm{E}[e^{it\hat{T}_E}] - e^{-t^2/2}\right| &= \left|\mathrm{E}[e^{it\hat{T}_E} - e^{-t^2/2}]\right| \\ &\leq \mathrm{E}\left[\left|e^{it\hat{T}_E} - e^{-t^2/2}\right|\right] \xrightarrow[n\to\infty]{} 0\end{aligned}$$

Hence we have that unconditionally, $\hat{T}_E$ converges to a $\mathcal{N}(0,1)$.

□

Proof. Corollary 1.

Firstly, by proposition 1.1 we have that for any realization of $\hat{S}$ denoted $S \in \mathcal{S}_{\text{strong}}$, one can build asymptotically valid *conditional* confidence intervals with coverage $(1-\alpha)$ in the usual way:

$$CI_\alpha(S) = \left[\hat{\tau}(S) - \frac{\sqrt{\hat{V}^{\hat{\tau}(S)}}}{\sqrt{n_E}} \cdot q_{1-\frac{\alpha}{2}} \ , \ \hat{\tau}(S) + \frac{\sqrt{\hat{V}^{\hat{\tau}(S)}}}{\sqrt{n_E}} \cdot q_{1-\frac{\alpha}{2}}\right]$$

where $\hat{V}^{\hat{\tau}(S)}$ is a consistent estimator of the asymptotic variance of $\hat{\tau}(S)$, and $q_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the $\mathcal{N}(0,1)$ distribution. Those CIs are asymptotically valid by proposition 1.1, i.e.:

$$\mathrm{P}[LATE \in CI_\alpha(\hat{S})] \mid \hat{S} = S \xrightarrow[n\to\infty]{} 1-\alpha$$

Now, by the law of iterated expectations, we have that:

$$\mathrm{P}[LATE \in CI_\alpha(\hat{S})] = E\left[E[\mathbb{1}\{LATE \in CI_\alpha(\hat{S})\}|\hat{S} = S]\right] \xrightarrow[n\to\infty]{} 1-\alpha$$

which is the second statement of corollary 1.

Now let us turn to the first statement, i.e.,

$$\lim_{n\to\infty} \mathrm{P}\left[\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha(S)] \leq \sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{TSLS}]\right] \;=\; 1$$

$\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha(S)]$ and $\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{TSLS}]$ are entirely governed by and strictly increasing in $\hat{V}^{\hat{\tau}(S)}$ and $\hat{V}^{\mathrm{TSLS}}$ respectively. Let $\hat{V}^{\hat{\tau}(S)}$ and $\hat{V}^{\mathrm{TSLS}}$ be estimators that converge in probability to $V^{\hat{\tau}(S)}$ and V^{TSLS}, and we assumed that S was such that we were in the inequality case of proposition 1.2, i.e.,

$$V^{\hat{\tau}_E(S)} < V^{\mathrm{TSLS}}$$

Let us denote by $\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^0(S)]$ and $\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{0,TSLS}]$ the (rescaled) CIs constructed with the true values of the asymptotic variances, i.e., $V^{\hat{\tau}(S)}$ and V^{TSLS} respectively. We thus have:

$$\forall \varepsilon_1 > 0,\ \lim_{n\to\infty} \mathrm{P}[\left|\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha(S) - \sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^0(S)\right| > \varepsilon] = 0$$

and

$$\forall \varepsilon_2 > 0,\ \lim_{n\to\infty} \mathrm{P}[\left|\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{TSLS}] - \sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{0,TSLS}]\right| > \varepsilon] = 0$$

Since $V^{\hat{\tau}_E(S)} < V^{\mathrm{TSLS}}$, we have that

$$\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^0(S) < \sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{0,TSLS}]$$

Hence we have:

$$\lim_{n\to\infty} \mathrm{P}\left[\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha(S)] < \sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{TSLS}]\right] = 1$$

□

Proof. Lemma 3.

A few elements need to be reminded to the reader in order to prove this lemma.

First of all, if we denote by $\hat{\tau}_1$ the estimator constructed using the fold $\mathcal{I}_2$ as the test sample and

$\mathcal{I}_1$ as the estimation sample, recall that we can decompose it as follows:

$$\sqrt{n_{(1)}}(\hat{\tau}_1 - LATE) = \frac{1}{\sqrt{n_{(1)}}} \sum_i \psi_{\hat{\tau}_1,i}$$

where the expression of the influence function is given by:

$$\begin{aligned}\psi_{\hat{\tau}_1,i} =& \frac{1}{p_{C,\hat{S}_{G,(2)}=1}} \left(\frac{1}{E[Z\hat{S}_{G,(2)}]} Z_i \hat{S}_{G_i,2} \cdot (\varepsilon_i - E[\varepsilon | Z=1, \hat{S}_{G,(2)}=1]) \right. \\ & \left. - \frac{1}{E[(1-Z)\hat{S}_{G,(2)}]} (1-Z_i) \hat{S}_{G_i,2} \cdot (\varepsilon_i - E[\varepsilon | Z=0, \hat{S}_{G,(2)}=1]) \right)\end{aligned}$$

with $\hat{S}_{g,(2)}$ denoting the selection indicator for group g computed in fold $\mathcal{I}_2$ as follows:

$$\hat{S}_{g,(2)} \equiv \mathbb{1}\left\{ \hat{\pi}^g_{(2)} > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right\}$$

where $n^g_{(2)}$ is the number of observations in group g in sample $\mathcal{I}_1$, $\hat{\pi}^g_{(2)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance of $\hat{\pi}^g_{(2)}$. Second, recall (from the proof of corollary 1 above) that:

$$\begin{aligned}\sqrt{n^g_{(2)}} \cdot \left[\hat{\pi}^g_{(2)} - \pi g \right] &= \frac{1}{\sqrt{n^g_{(2)}}} \cdot \left[\sum_{i=1}^{n^g_{(2)}} \underbrace{\left(\frac{Z_i \left(D_i - \mathbb{E}[D \mid Z=1] \right.}{E[Z]} + \frac{(1-Z_i) \cdot (D_i - \mathbb{E}[D \mid Z=0])}{1 - \mathbb{E}[2]} \right)}_{\equiv \tilde{\psi}^g_i} \right] \\ &= \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i\end{aligned}$$

The above formulas make it clear that the potential source of dependence between $\hat{\tau}_1$ and $\hat{\tau}_2$ lies in $\hat{S}_{g,(2)}$, that appears in the influence function of $\hat{\tau}_1$ and is computed based on observations from fold $\mathcal{I}_2$, also used in $\hat{\tau}_2$. We will now study the (asymptotic) dependence of $\hat{S}_{g,(2)}$ on $\tilde{\psi}^g_n$, the n^{th} individual influence function entering in $\hat{\pi}^g_{(2)}$. For groups g such that $\pi^g > 0$ (strong first-stage), we have that $\mathrm{P}[\hat{S}_{g,(2)} = 1] \xrightarrow[n\to\infty]{} 1$ and $\hat{S}_{g,(2)}$ becomes essentially deterministic, hence

asymptotically there aren't any dependence issues for such groups. We will therefore focus on groups g such that $\pi^g = 0$. For any such group g, and for a given number of observations $n^g_{(2)}$ in this group (in fold $\mathcal{I}_2$), we have:

$$\begin{aligned}
\hat{S}^{(n^g_{(2)})}_{g,(2)} &= \mathbb{1}\left\{ \hat{\pi}^g_{(2)} > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right\} \\
&= \mathbb{1}\left\{ \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right\} \\
&= \mathbb{1}\left\{ F^{g,n^g_{(2)}} > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right\}
\end{aligned}$$

where we defined: $F^{g,n^g_{(2)}} \equiv \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i$. Hence we can study the probability that any additional observation modifies the value of $\hat{S}^{(n^g_{(2)})}_{g,(2)}$ as follows:

$$\begin{aligned}
&\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g,(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g,(2)} = 1 \right] \\
&= \mathrm{P}\left[F^{g,n^g_{(2)}-1} \leq \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}-1}} \cdot (q_{1-\alpha} - \epsilon)\,, \quad F^{g,n^g_{(2)}} > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right] \\
&\leq \mathrm{P}\left[\left| F^{g,n^g_{(2)}-1} \right| \leq \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}-1}} \cdot (q_{1-\alpha} - \epsilon)\,, \quad \left| F^{g,n^g_{(2)}} \right| > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\alpha} \right] \\
&= \mathrm{P}\left[\left| \left(n^g_{(2)} - 1 \right) \cdot F^{g,n^g_{(2)}-1} \right| \leq \sqrt{n^g_{(2)}-1} \cdot \hat{\sigma}^{\pi^g} \cdot (q_{1-\alpha} - \epsilon)\,, \quad n^g_{(2)} \cdot \left| F^{g,n^g_{(2)}} \right| > \sqrt{n^g_{(2)}} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\alpha} \right]
\end{aligned}$$

Notice that:

$$\begin{aligned}
n^g_{(2)} \cdot \left| F^{g,n^g_{(2)}} \right| &= \left| \tilde{\psi}^g_{n^g_{(2)}} + \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}-1} \tilde{\psi}^g_i \right| \\
&= \left| \tilde{\psi}^g_{n^g_{(2)}} + \left(n^g_{(2)} - 1 \right) \cdot F^{g,n^g_{(2)}-1} \right| \\
&\leq \left| \tilde{\psi}^g_{n^g_{(2)}} \right| + \left(n^g_{(2)} - 1 \right) \cdot \left| F^{g,n^g_{(2)}-1} \right| \qquad \text{(by the triangle inequality)}
\end{aligned}$$

where $\tilde{\psi}^g_{n^g_{(2)}}$ denotes the influence function of the $n^g_{(2)}$-th observation. Hence we get:

$$
\begin{aligned}
&\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g,(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g,(2)} = 1\right] \\
&\leq \mathrm{P}\left[\left|\left(n^g_{(2)}-1\right)\cdot F^{g,n^g_{(2)}-1}\right| \leq \sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\cdot(q_{1-\alpha}-\epsilon),\quad n^g_{(2)}\cdot\left|F^{g,n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha}\right] \\
&= \mathrm{P}\left[\left|\left(n^g_{(2)}-1\right)\cdot F^{g,n^g_{(2)}-1}\right| \leq \sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\cdot(q_{1-\alpha}-\epsilon),\quad n^g_{(2)}\cdot\left|F^{g,n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha}\right] \\
&\leq \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha} - \left(n^g_{(2)}-1\right)\cdot\left|F^{g,n^g_{(2)}-1}\right|,\quad \left(n^g_{(2)}-1\right)\cdot\left|F^{g,n^g_{(2)}-1}\right| \leq \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot(q_{1-\alpha}-\epsilon)\right] \\
&\leq \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha} - \left(n^g_{(2)}-1\right)\cdot\left|F^{g,n^g_{(2)}-1}\right|,\right. \\
&\qquad \left.\left(n^g_{(2)}-1\right)\cdot\left|F^{g,n^g_{(2)}-1}\right| \leq \sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\cdot(q_{1-\alpha}-\epsilon)\right] \\
&\leq \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\cdot\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha} - \sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\cdot(q_{1-\alpha}-\epsilon)\right] \\
&= \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \hat{\sigma}^{\pi^g}\cdot q_{1-\alpha}\cdot\left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\cdot\sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\right]
\end{aligned}
$$

For $n^g_{(2)}$ large enough, we have:

$$
\hat{\sigma}^{\pi^g}\cdot q_{1-\alpha}\cdot\left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\cdot\sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g} \approx \epsilon\cdot\sqrt{n^g_{(2)}-1}\cdot\sigma^{\pi^g} \xrightarrow[n\to\infty]{} \infty
$$

Hence we get:

$$
\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g,(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g,(2)} = 1\right] \leq \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \hat{\sigma}^{\pi^g}\cdot q_{1-\alpha}\cdot\left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\cdot\sqrt{n^g_{(2)}-1}\cdot\hat{\sigma}^{\pi^g}\right] \xrightarrow[n\to\infty]{} 0
$$

Therefore, for n (and therefore $n^g_{(2)}$) large enough, $\hat{S}^{(n^g_{(2)})}_{g,(2)}$ becomes independent of any single observations from sample $\mathcal{I}_2$, and consequently so does $\hat{\tau}_1$. Therefore, under those standard asymptotics, $\hat{\tau}_1$ and $\hat{\tau}_2$ are asymptotically independent. □

Proof. Proposition 2.

Lemma 4 states that as n_T goes to infinity, there are only a certain set of values that $\hat{S}$ can take, denoted $\mathcal{S}_{\text{strong}}$. When S takes its value in some subsets of $\mathcal{S}_{\text{strong}}$, the analysis of the asymptotic distribution of $\hat{\tau}(S)$ is rather straightforward. Indeed, as long as all groups with weak first-stages are included in the selected sample, we are back to the case previously studied in proposition 1

as we can recast the problem as one with two groups:

1. One including all groups with a strong or a weak first-stage, plus groups with zero first stages that are included in the selected sample defined by S. By construction, overall this group has a strong first-stage.
2. One including all groups with zero first-stages that are not included in the selected sample defined by S. By construction, overall this group has a zero first-stage.

Then we know by proposition 1 that the asymptotic distribution of $\hat{\tau}(S)$ in such a setting will be centered on the LATE. Formally, let us defined:

$$\mathcal{S}^0_{\text{strong}} \equiv \{S \in \mathcal{S}_{\text{strong}} \ : \ \forall g \in \mathcal{G}_W, S_g = 1\}$$
$$\mathcal{S}^1_{\text{strong}} \equiv \{S \in \mathcal{S}_{\text{strong}} \ : \ \exists g \in \mathcal{G}_W, S_g = 0\}$$

By proposition 1 and the argument above, we have:

$$\forall S \in \mathcal{S}^0_{\text{strong}}, \quad \sqrt{n_E} \cdot (\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(0, V^S)$$

Now, we turn to the case where S belongs to the set $\mathcal{S}^1_{\text{strong}}$. This includes all cases in which some of the groups with a weak share of compliers get excluded from the restricted sample. We can always reframe such a situation by redefining two groups:

1. Group 1 including all selected groups as defined S. By construction, overall this group has a strong first-stage.
2. Group 2 including all excluded groups. By construction, since (by definition of $\mathcal{S}^1_{\text{strong}}$) it contains groups with weak first-stages, overall this group has a weak first-stage as well.

Recasting the problem in this way places it in the setting studied in lemma 8, which proves the result.

□

Proof. Theorem 1.

Theorem 1.1 Lemma 8 and proposition 2 show that for all possible values of the selection vector S in $\mathcal{S}_{\text{strong}}$ — that is, all the values that the random vector $\hat{S}$ (determined in sample $\mathcal{I}_T$) takes with non-zero probability asymptotically — the asymptotic bias of $\sqrt{n_E}\,(\hat{\tau}(S) - LATE)$ is of the form:

$$C \cdot \left(LATE^{\mathcal{G}_W^S} - LATE^{\mathcal{G}_S} \right)$$

where C denotes a finite constant, $LATE^{\mathcal{G}_W^S}$ denotes the LATE among groups with a weak first-stage that are selected according to S, and $LATE^{\mathcal{G}_S}$ denotes the LATE among groups with a strong first-stage (always selected for $S \in \mathcal{S}_{\text{strong}}$). A sufficient condition for this asymptotic bias to be negligible is assumption 4, that implies: $LATE^{\mathcal{G}_W^S} - LATE^{\mathcal{G}_S} = o(1)$. Under this assumption, we have:

$$\forall S \in \mathcal{S}_{\text{strong}}, \quad B(S) = 0$$

Hence the first result. Notice further that under assumption 4, groups $g \in \mathcal{G}_{WIV}$ can be treated essentially in the same way as groups $g \in \mathcal{G}_0$. Indeed, one can redefine the target estimand as $LATE + B(S)$ — which is first-order equivalent to $LATE$ under assumption 4 — and the influence function of $\hat{\tau}(S)$ has naturally the same form as the one studied in 1. Hence following the reasoning of the proofs of proposition 1.1 and 1.2 — yet using appropriate central limit theorem for triangular arrays (Lindeberg-Feller CLT) instead of the standard CLT — we get:

$$V^{\hat{\tau}(S)} \leq V^{TSLS}$$

Theorem 1.2 The proof follows the exact same line of reasoning as in the proof of 1.3, yet making use of assumption 4 and its implication in theorem 1.1 to get the result. Indeed, the proof relies on the consistency of $\hat{\tau}(S)$ for any $S \in \mathcal{S}_{strong}$, which (in the presence of groups with weak first-stages) is guaranteed under assumption 4 as shown above in the proof of theorem 1.1.

□

A.2 Proofs of useful lemmas

Proof. Lemma 2.

The random vector $\hat{S}$ stacks the tests statistics:

$$T^g_{\alpha,n_T} = \mathbb{1}\left\{\sqrt{n^g_T} \cdot \frac{\hat{\pi}^g}{\hat{\sigma}^g} > q_{1-\alpha}\right\}$$

where n^g_T denotes the test sample size in group $X = x$. Notice that here we are assuming that the sample sizes of the groups are not random, which is asymptotically equivalent to sampling with a fixed fraction. We also denote by $\hat{\pi}^g$ the estimator of π^g, $\hat{\sigma}^g$ the estimator of the variance of $\hat{\pi}^g$, and $q_{1-\frac{\alpha}{2}}$ the $1-\frac{\alpha}{2}$ quantile of a $\mathcal{N}(0,1)$.

The t-test being consistent against any alternative well separated from 0, we have:

$$\forall g \in \mathcal{G}_S, \quad \lim_{n_T \to \infty} \Pr[T^g_{\alpha,n_T} = 1] = 1$$

since we have: $\forall g \in \mathcal{G}_S,\ \pi^g > 0$.

As the level of the test is α, we also have:

$$\forall g \in \mathcal{G}_0, \quad \lim_{n_T \to \infty} \Pr[T^g_{\alpha,n_T} = 1] = \alpha$$

since we have: $\forall g \in \mathcal{G}_0,\ \pi^g = 0$. □

Proof. Lemma 4.

The proof follows exactly the same steps as for lemma 2 for groups with 0 and strong first-stages, yet using appropriate central limit theorem for triangular arrays (Lindeberg-Feller CLT) instead of the standard CLT — as the presence of groups with weak first-stages requires that the DGP changes with n. For groups with weak first-stages, we have that the first-stage parameter takes the form $\pi^g = \frac{H^g}{\sqrt{n_T}}$ — where H^g is what if often called the "location parameter". Therefore, we have:

$$\sqrt{n_T} \cdot \frac{\hat{\pi}^g}{\hat{\sigma}^g} \xrightarrow{d} \mathcal{N}(H^g, 1)$$

The quantiles of a $|\mathcal{N}(b,1)|$ are increasing in b, and by assumption $H^g > 0$. Hence using the same definition of the test statistic as in the proof of lemma 2, we get:

$$\forall g \in \mathcal{G}_W, \quad \lim_{n_T \to \infty} \Pr[T^g_{\alpha,n_T} = 1] > \alpha$$

□

LEMMA 5 (Influence function of the ratio of two asymptotically linear estimators). *Let $\widehat{A}$ and $\widehat{B}$ be asymptotically linear estimators:*

$$\sqrt{n}(\widehat{A} - A) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} a_i + o_P(1)$$

and

$$\sqrt{n}(\widehat{B} - B) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_i + o_P(1)$$

with $\mathrm{E}[a_i] = \mathrm{E}[b_i] = 0$. Then we have:

$$\sqrt{n}\left(\frac{\widehat{A}}{\widehat{B}} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_i - (A/B)b_i}{B} + o_P(1)$$

Proof. There is a general relationship which is easy to verify:

$$\frac{\hat{A}}{\hat{B}} - \frac{A}{B} = \left(\frac{\hat{A} - A}{B} - \frac{A}{B}\frac{\hat{B} - B}{B}\right) \cdot \left(1 - \frac{\hat{B} - B}{\hat{B}}\right)$$

Plugging in the asymptotically linear formula into the first formula, we obtain:

$$\begin{aligned}
\sqrt{n}\left(\frac{\hat{A}}{\hat{B}} - \frac{A}{B}\right) &= \left(\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} a_i + o_P(1)}{B} - \frac{A}{B}\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} b_i + o_P(1)}{B}\right) \cdot \left(1 - \frac{\hat{B} - B}{\hat{B}}\right) \\
&= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_i - (A/B)b_i}{B} + o_P(1)\right) \cdot \left(1 - \frac{o_P(1)}{O_P(1)}\right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_i - (A/B)b_i}{B} + o_P(1)
\end{aligned}$$

where we went from the first to the second equality because (i) $(\hat{B} - B) = o_P(1)$ by the weak

LLN, since it is an empirical mean of terms b_i with expectation 0, (ii) $\hat{B} = O_p(1)$ since it converges in probability to $B < \infty$, and (iii) since $O_P(1)^{-1} = O_P(1)$ and $o_P(1) \cdot O_P(1) = o_P(1)$, we have: $\frac{\hat{B}-B}{\hat{B}} = o_P(1)$.

□

LEMMA 6 (Influence function of the estimator of a CEF). *The influence function of the estimator* $\frac{\sum_i Z_i Y_i}{\sum Z_i}$ *of the conditional expectation function* $E[Y|Z=1]$ *is given by:* $\psi_i = \frac{Z_i(Y_i - \mathrm{E}[Y|Z=1])}{\mathrm{E}[Z]}$.

Proof.

$$
\begin{aligned}
&\sqrt{n}\left(\frac{\sum_i Z_i Y_i}{\sum Z_i} - E[Y|Z=1]\right) \\
&= \sqrt{n}\left(\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{\sum Z_i}\right) \\
&= \sqrt{n}\left(\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{E[Z]}\right)\frac{E[Z]}{\sum Z_i} \\
&= \frac{1}{\sqrt{n}}\left(\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{E[Z]}\right)\frac{E[Z]}{\frac{\sum Z_i}{n}} \\
&= \frac{1}{\sqrt{n}}\left(\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{E[Z]}\right) + \frac{1}{\sqrt{n}}\left(\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{E[Z]}\right)\frac{E[Z] - \frac{\sum Z_i}{n}}{\frac{\sum Z_i}{n}} \\
&= \frac{1}{\sqrt{n}}\frac{\sum_i Z_i (Y_i - E[Y|Z=1])}{E[Z]} + o_P(1)
\end{aligned}
$$

or equivalently from lemma 5, which gives the same influence function when setting $\hat{A} = \sum_i Z_i Y_i$, $A = \mathrm{E}[ZY] = \mathrm{E}[Y|Z=1]\,\mathrm{E}[Z]$, $a_i = Z_i Y_i - \mathrm{E}[Y|Z=1]\,\mathrm{E}[Z]$, and $\hat{B} = \sum_i Z_i$, $B = \mathrm{E}[Z]$, $b_i = Z_i - \mathrm{E}[Z]$. □

LEMMA 7 (Asymptotic distribution of 2SLS/Wald estimator).

$$\sqrt{n}(\hat{\tau}^{Wald} - LATE) \xrightarrow{d} \mathcal{N}(0, V(\psi_{\hat{\tau}^{Wald},i}))$$

where $V(\psi_{\hat{\tau}^{Wald},i})$ *equals:*

$$V(\psi_{\hat{\tau}^{Wald},i}) = \frac{1}{p_C^2}\left(\frac{1}{p}V[\varepsilon|Z=1] + \frac{1}{1-p}V[\varepsilon|Z=0]\right)$$

Proof. The Wald estimator is merely a ratio of difference of conditional expectation function

(CEF) estimators — and it estimates the LATE, which is a ratio of difference of CEFs. Therefore, we can see it as the combination of several asymptotically linear estimators:

$$\begin{aligned}
\hat{A} &= \left(\sum_i Z_i\right)^{-1} \sum_i Z_i Y_i, & A &= E[Y|Z=1] \\
\hat{B} &= \left(\sum_i (1-Z_i)\right)^{-1} \sum_i (1-Z_i) Y_i, & B &= E[Y|Z=0] \\
\hat{C} &= \left(\sum_i Z_i\right)^{-1} \sum_i Z_i D_i, & C &= E[D|Z=1] \\
\hat{D} &= \left(\sum_i (1-Z_i)\right)^{-1} \sum_i (1-Z_i) D_i, & D &= E[D|Z=0] \\
&\Rightarrow LATE = \frac{A-B}{C-D} \\
&\Rightarrow \widehat{\tau}^{Wald} = \frac{\hat{A}-\hat{B}}{\hat{C}-\hat{D}}
\end{aligned}$$

By lemma 6, the influence functions of $\hat{A}, \hat{B}, \hat{C}$ and $\hat{D}$ are given respectively by:

$$\begin{aligned}
a_i &= \frac{Z_i(Y_i - E[Y|Z=1])}{E[Z]} \\
b_i &= \frac{(1-Z_i)(Y_i - E[Y|Z=0])}{1-E[Z]} \\
c_i &= \frac{Z_i(D_i - E[D|Z=1])}{E[Z]} \\
d_i &= \frac{(1-Z_i)(Y_i - E[D|Z=0])}{1-E[Z]}
\end{aligned}$$

We then have:

$$\sqrt{n}(\widehat{\tau}^{Wald} - LATE) = \frac{1}{\sqrt{n}} \sum_i \psi_{\widehat{\tau}^{Wald},i} + o_P(1)$$

where (following lemma 5) $\psi_{\hat{\tau}^{Wald},i}$ is given by:

$$\begin{aligned}
\psi_{\hat{\tau}^{Wald},i} &= \frac{(a_i - b_i) - LATE \cdot (c_i - d_i)}{C - D} \\
&= \frac{1}{\pi}\left(\frac{Z_i(Y_i - E[Y|Z=1])}{E[Z]} - \frac{(1-Z_i)(Y_i - E[Y|Z=0])}{1-E[Z]}\right. \\
&\qquad \left. -LATE \cdot \left(\frac{Z_i(D_i - E[D|Z=1])}{E[Z]} - \frac{(1-Z_i)(D_i - E[D|Z=0])}{1-E[Z]}\right)\right) \\
&= \frac{1}{\pi}\left(\frac{1}{p}Z_i \cdot (\varepsilon_i - E(\varepsilon|Z=1)) - \frac{1}{1-p}(1-Z_i)\cdot(\varepsilon_i - E(\varepsilon|Z=0))\right)
\end{aligned}$$

where $\varepsilon = Y - LATE \cdot D$ is the structural error term of the second stage, and $\pi = \mathrm{E}[D(1) - D(0)]$ is the share of compliers. As expected from an influence function, one can check that $E[\psi_{\hat{\tau}^{Wald},i}] = 0$. It follows that asymptotically,

$$\sqrt{n}(\hat{\tau}^{Wald} - LATE) \xrightarrow{d} \mathcal{N}(0, V(\psi_{\hat{\tau}^{Wald},i}))$$

where $V(\psi_{\hat{\tau}^{Wald},i})$ equals:

$$\begin{aligned}
V(\psi_{\hat{\tau}^{Wald},i}) &= E(\psi^2_{\hat{\tau}^{Wald},i}) \\
&= E(\psi^2_{\hat{\tau}^{Wald},i}|Z=1)p + E(\psi^2_{\hat{\tau}^{Wald},i}|Z=0)(1-p) \\
&= \frac{1}{\pi^2}\left(\frac{1}{p}E[(\varepsilon - E[\varepsilon|Z=1])^2|Z=1] + \frac{1}{1-p}E[(\varepsilon - E[\varepsilon|Z=0])^2|Z=0]\right) \\
&= \frac{1}{\pi^2}\left(\frac{1}{p}V[\varepsilon|Z=1] + \frac{1}{1-p}V[\varepsilon|Z=0]\right)
\end{aligned}$$

□

LEMMA 8 (Bias of the test-and-select estimator in the 3-group case). *Let's consider a case with only three groups: a group with a strong first-stage ($\pi^1 > 0$), a group with a weak first-stage ($\pi^2 = H^2/\sqrt{n}$), and a group with a zero first-stage ($\pi^3 = 0$). Under assumption 3, and we have:*

$$\sqrt{n_E}\,(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(B(S), V^S))$$

with $B(S) = \frac{H^2 \cdot \Pr[G=2]}{\pi} \cdot (LATE^1 - LATE^2)$ if group 2 is not selected.

Proof. Let's consider a case with only three groups: a group with a strong first-stage ($\pi^1 > 0$), a group with a weak first-stage ($\pi^2 = H^2/\sqrt{n}$), and a group with a zero first-stage ($\pi^3 = 0$).
Group 1 is always selected as asymptotically (as n_T goes to infinity), the selection procedure selects groups with a strong first-stage with probability 1.
Group 3 being selected or not does not affect the expectation of the limiting distribution of the ($\sqrt{n}-$scaled) resulting estimator, as shown in the proof of proposition 1.1. Hence we can ignore group 3 — or simply redefine group 1 or group 2 as including group 3 as well — without any changes in the result, and simply consider the two following cases:

1. Group 1 is selected, group 2 is selected
2. Group 1 is selected, group 2 is not selected

In the first case, the resulting estimator is the standard Wald estimator[46] computed on the whole estimation sample... hence it is $\sqrt{n}-$consistent (no asymptotic bias).
In the second case, the resulting estimator corresponds to the Wald estimator computed on group 1. Hence it is a $\sqrt{n}-$consistent estimator for the LATE conditional on being in group 1, which we define below:

$$LATE^1 \equiv \mathrm{E}[Y(1) - Y(0)|D(1) > D(0), G = 1]$$

In other words, denoting by $\hat{\tau}(S^2)$ the estimator in case 2, we have:

$$\sqrt{n_E} \cdot \left(\hat{\tau}(S^2) - LATE^1\right) \xrightarrow{d} \mathcal{N}(0, V^{S^2})$$

Now, since we are interested in the limiting distribution of $\sqrt{n_E} \cdot (\hat{\tau}(S^2) - LATE)$, what is left to study is the behavior of:

$$\sqrt{n_E} \cdot \left(LATE^1 - LATE\right) \xrightarrow{?} 0$$

At first, the quantities involved above might seem independent of n_E. The dependence of $LATE$ on n_E comes from the fact that the share of compliers in group 2 depends on n_E, as we

[46] Whether or not group 3 (group with no first-stage at all) is included or not in the estimation will have an effect on the variance of the resulting estimator, as argued in the first part of this paper (with standard asymptotics).

have: $\pi^2 = H^2/\sqrt{n_E}$.

We have:

$$
\begin{aligned}
LATE^g &= \mathrm{E}[Y(1) - Y(0) | D(1) > D(0), G = g] \\
LATE &= \mathrm{E}[Y(1) - Y(0) | D(1) > D(0)] \\
&= \mathrm{E}[Y(1) - Y(0) | D(1) > D(0), G = 1] \cdot \Pr[G = 1 | D(1) > D(0)] \\
&\quad + \mathrm{E}[Y(1) - Y(0) | D(1) > D(0), G = 2] \cdot \Pr[G = 2 | D(1) > D(0)] \quad \text{(Law of iterated exp.)} \\
&= LATE^1 \cdot \frac{\Pr[D(1) > D(0) | G = 1] \cdot \Pr[G = 1]}{\Pr[D(1) > D(0)]} \\
&\quad + LATE^2 \cdot \frac{\Pr[D(1) > D(0) | G = 2] \cdot \Pr[G = 0]}{\Pr[D(1) > D(0)]} \quad \text{(Bayes' rule)} \\
&= LATE^1 \cdot \frac{\pi^1 \cdot \Pr[G = 1]}{\pi} + LATE^2 \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi}
\end{aligned}
$$

where the last line uses our standard notations:

$$
\begin{aligned}
\pi^g &\equiv \mathrm{E}[D(1) - D(0) | G = g] \\
\pi &\equiv \mathrm{E}[D(1) - D(0)] = \pi^1 \cdot \Pr[G = 1] + \pi^2 \cdot \Pr[G = 2]
\end{aligned}
$$

Hence we get:

$$
\begin{aligned}
\sqrt{n_E} \cdot \big(LATE^1 - LATE\big) &= \sqrt{n_E} \cdot \left(LATE^1 \cdot \left(1 - \frac{\pi^1 \cdot \Pr[G = 1]}{\pi} \right) - LATE^2 \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi} \right) \\
&= \sqrt{n_E} \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi} \cdot \big(LATE^1 - LATE^2\big) \\
&= \frac{H^2 \cdot \Pr[G = 2]}{\pi} \cdot \big(LATE^1 - LATE^2\big)
\end{aligned}
$$

Therefore, we have in this case:

$$
\begin{aligned}
\sqrt{n_E} \cdot \big(\hat{\tau}(S^2) - LATE\big) &= \sqrt{n_E} \cdot \big(\hat{\tau}(S^2) - LATE^1\big) + \sqrt{n_E} \cdot \big(LATE^1 - LATE\big) \\
&= \sqrt{n_E} \cdot \big(\hat{\tau}(S^2) - LATE^1\big) + B(S^2) \\
&\xrightarrow{d} \mathcal{N}(B(S^2), V^{S^2}) \quad \text{(Slutsky's lemma)}
\end{aligned}
$$

with $B(S^2) \equiv \frac{H^2 \cdot \Pr[G=2]}{\pi} \cdot (LATE^1 - LATE^2)$.

□

LEMMA 9 (Bias of Coussens and Spiess (2021) estimator). *Under assumption 4, the estimator studied in Coussens and Spiess (2021) has a first-order bias.*

Proof. The proof follows the one of Proposition 6 in Coussens and Spiess (2021). The only difference resides in the fact that assumption 4 does not assume that all treatment effects are of order $1/\sqrt{n}$, but simply that the treatment effect heterogeneity is. We will use Coussens and Spiess (2021) notations.

Assumption 4, translated in their notations, can be written as: $\tau(X) = \lambda + \frac{\mu(X)}{\sqrt{n}}$.

Their proof goes as follows:

$$\sqrt{n}\left(\hat{\tau}_w - \tau\right) = \sqrt{n}\left(\hat{\tau}_w - \tau_w\right) + \underbrace{\sqrt{n}\left(\tau_w - \tau\right)}_{=B_w} = \sqrt{n}\left(\hat{\tau}_w - \tau_w\right) + B_w \xrightarrow{d} \mathcal{N}\left(B_w, V_w\right)$$

where:

$$B_w = \frac{\text{Cov}(\mu(X), w(X) \mid D(1) > D(0))}{\text{E}[w(X) \mid D(1) > D(0)]}$$

The convergence of $\sqrt{n}\left(\hat{\tau}_w - \tau_w\right)$ to a normal centered on 0 results from proposition 5 in Coussens and Spiess (2021). τ_w is the estimand towards which their estimator $\hat{\tau}_w$ converges in the absence of any restrictions on heterogeneity, and τ is the LATE parameter.

We simply need to study whether we still have:

$$\sqrt{n}\left(\tau_w - \tau\right) = B_w$$

under the treatment effect modeling $\tau(X) = \lambda + \frac{\mu(X)}{\sqrt{n}}$.

Indeed, we have:

$$
\begin{aligned}
&\sqrt{n}\left(\tau_w - \tau\right) \\
&= \frac{\mathrm{E}[\alpha(X)w(X)\sqrt{n}\tau(X)]}{\mathrm{E}[\alpha(X)w(X)]} - \frac{\mathrm{E}[\alpha(X)\sqrt{n}\tau(X)]}{\mathrm{E}[\alpha(X)]} \\
&= \frac{\mathrm{E}[\alpha(X)w(X)\mu(X)]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)\mu(X)]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]} \\
&\quad - \frac{\mathrm{E}[\alpha(X)w(X)\sqrt{n}\mu]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)\sqrt{n}\mu]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]} \\
&= \frac{\frac{\mathrm{E}[\alpha(X)w(X)\mu(X)]}{\mathrm{E}[\alpha(X)]} - \frac{\mathrm{E}[\alpha(X)\mu(X)]}{\mathrm{E}[\alpha(X)]}\frac{\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]}}{\frac{\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]}} - \sqrt{n}\mu \underbrace{\frac{\mathrm{E}[\alpha(X)w(X)]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}}_{=0} \\
&= B_w + 0
\end{aligned}
$$

Hence the result of proposition 6 of Coussens and Spiess (2021) remains under our own assumption 4 on treatment effect heterogeneity. □

A.3 Additional simulations

Illustrating the necessity of data-splitting

This DGP has been selected in order to illustrate the bias of a naïve selection rules that would test first-stages, select groups accordingly, and estimate the LATE in the same sample without any sample split. Indeed, this DGP does not feature any first-stage heterogeneity nor treatment effect heterogeneity across groups — hence the potential bias of any selection procedure would be of a different nature than the one studied in the simulations presented in section 5. Yet as discussed in the end of section 2, pre-testing on the first-stage might generate bias in the estimator of the first-stage coefficient (see lemma 1) and thus ultimately in the resulting LATE estimator.
The DGP parameters are the following:

$$
\text{DGP0} \equiv \left(N = 1000, J \in \{10, 20, 30, 50\}, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\delta\varepsilon} = 0.3, \sigma_\eta = 1, \alpha = 0.0 \right)
$$

We vary the number of groups as a way to exacerbate the pre-testing issue in simulations. We report in Table 1.1 below the results of a Monte-Carlo simulation following DGP0 (with a number of groups $J = 30$) with 10,000 repetitions. In summary, this DGP generates a sample of size $N = 1000$, divided randomly into 30 groups (i.e., roughly 33 observations per group). The share of compliers in the sample (and thus in each randomly created group on average) is 25%. In such a setting, we do not expect our procedure to yield any gains, as there are no sub-populations without compliers. Yet selecting "naïvely" based on a t-test — without any sample split to alleviate the pre-testing issues mentioned above — might introduce a bias in the estimation of the LATE, that could invalidate the inference conducted based on such estimator. In order to answer this question, Table 1.1 reports the bias and coverage rate of 95%-confidence intervals of three estimators of the LATE over 10,000 Monte-Carlo repetitions. The first column reports the performance of the 2SLS estimator, the second column the performance of our proposed methodology *with* sample splitting and cross-fitting, and the third column a "naïve" version of our methodology that would test, select and estimate the LATE in the same sample without any sample split. The results show that the naïve version of the Test-and-Select estimator exhibits a clear bias (-0.221), which is ultimately detrimental to the coverage of its associated $95\%-$confidence interval that fail to cover at their nominal rate (0.861). Our proposed methodology that associates the Test-and-Select procedure with sample splitting and (2-folds) cross-fitting yields a much less biased estimator (0.097), and valid coverage (0.976). The remaining bias despite the use of data splitting and cross-fitting could be explained by the finite sample bias of 2SLS estimator.[47]

Figure 1.3 reports the bias of the three estimators presented in Table 1.1 for a varying number of groups. As can be seen in this graph, the bias generated by pre-testing and estimating in the same sample is larger when the number of observations per group is lower (larger number of groups). Yet our sample-splitting strategy corrects this bias equally well no matter the number of groups considered.

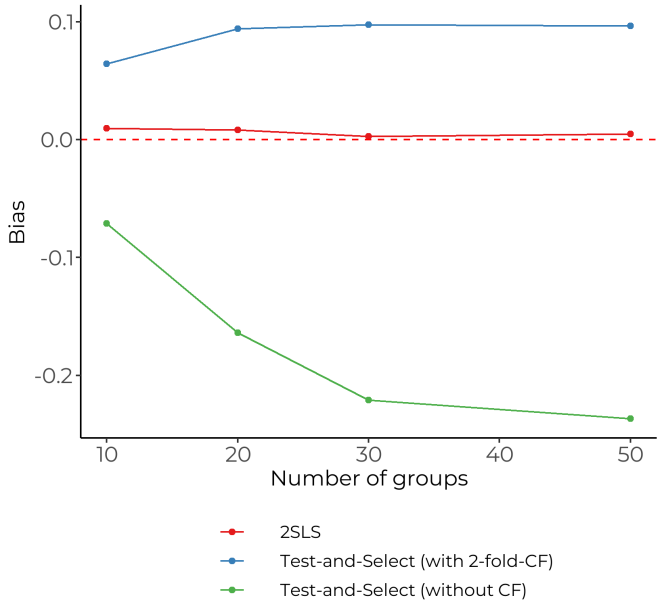
[47]Indeed, ultimately our Test-and-Select procedure with cross-fitting estimates the LATE by 2SLS on a smaller sample than the standard 2SLS estimator presented in the first column of Table 1.1. Therefore, its larger bias (0.097 vs. 0.003) could be explained by the finite sample bias of the 2SLS estimator, that vanishes as the sample size used for estimation grows.

Table 1.7 – Pre-test bias, and the use of cross-fitting

	2SLS	Test-and-Select (with 2-fold-CF)	Test-and-select (without CF)
Bias	0.003	0.097	-0.221
Coverage	0.953	0.976	0.861

Notes: This table presents the results of a simulation using the DGP0 described in section 5, with a number of groups of 30 — i.e., around 33 observations per group. In rows, we report the bias (with respect to the LATE parameter) and the coverage rate of 95%-confidence intervals. The first column reports the performance of the 2SLS estimator, the second column the performance of our proposed methodology *with* sample splitting and cross-fitting, and the third column a "naïve" version of our methodology that would test, select and estimate the LATE in the same sample without any sample split.

Figure 1.3 – Bias from lack of data-splitting as function of the number of groups



Notes: This figure shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP0, described in the text. Three different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select estimator with cross-fitting using 2 folds in blue, and a version of our Test-and-Select without data-splitting nor cross-fitting in green.

Chapter 2

Stigma and Benefit Take-up: Evidence from English Tabloid Newspapers

*Joint work with **Emily Silcock***

1 Introduction

Media has often been accused of stigmatising certain groups in society. This has the potential to affect their real-world economic behaviour, if making a certain decision risks increasing this stigma. In this paper, we focus on one stigmatised group - benefit recipients - and ask whether stigmatising media affects the decision of eligible people to take up social benefits.

Non-take-up is a pervasive problem, with take-up of some social benefits estimated to be lower than 30% in some European countries (Dubois et al., eds, 2015). Previous work shows that poorer and more marginalised individuals are less likely to take up benefits (Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2019), making non-take-up a serious impediment to the targetting of social assistance. However, while a wealth of qualitative and survey evidence attests to the presence of stigma around social benefits (Baumberg, 2016; Morrison, 2019), there is very little causal evidence on whether stigma actually affects take-up.

Given this gap in knowledge, this paper investigates the effect of media stigmatisation of benefit recipients on benefit take-up.

We focus on newspaper articles that contain negative content about benefit recipients from the British[1] tabloid newspaper, *The Sun*. During the period of study, this was the paper with the highest circulation in the UK, as well as arguably the one with the most negative coverage of benefit recipients.[2] To collect these articles, we trained a Deep Neural Network to classify whether articles are about social benefits or not. We use this to extract all *Sun* articles over the period 2013-2019 which are on this topic.[3] We then manually labelled which of these articles contained explicit negative sentiments towards people who are receiving or seeking benefits. Articles tend to cluster together in time, often with a news article on one day followed by commentary on the same or subsequent days. We refer to these clusters as 'stories'.

We provide causal evidence that these stories led to a decrease in the take-up of social benefits. We match daily data on the number of applications to Universal Credit, a composite working-age means-tested benefit, with the dates of the first article in a story containing negative coverage of benefit recipients. Our empirical strategy exploits the plausible local exogeneity of the timing of publication of newspaper articles. The publication of an article about benefit recipients is unlikely to be independent of long run take-up trends. However, we argue that it is exogenous to the number of people applying for benefits on a given day, compared to the day before. Using an event study approach, we examine whether applications to Universal Credit are affected by the publication of a story containing negative coverage of benefit recipients.

We find a large and significant decrease in the number of applications to Universal Credit in the days following such a story. On average, the number of applications decreases by 4% to 5% in the three days after publication. These results are driven by geographical regions with lower incomes and higher benefit application rates. We also find suggestive evidence that the results are driven by stories positioned nearer to the front of the newspaper, as determined by page number. We find no significant difference in the number of applications following a story

[1]The reader will recall that Great Britain (roughly) comprises England, Scotland and Wales, while the United Kingdom also includes Northern Ireland. In this paper we focus purely on England, due to data availability.

[2]*The Sun* was found to have the most negative coverage of benefit recipients of all British newspapers in a report by Turn2us, part of anti-poverty charity Elizabeth Finn, in 2012. `https://www.theguardian.com/news/datablog/2012/nov/20/benefits-stigma-newspapers-report-welfare`.

[3]For 2013-2016 abstracts of articles are available in ProQuest which allows us to run this classifier. After 2017 ProQuest stopped processing abstracts and only titles are available. Therefore we use keyword searches on LexisNexis (which does not allow us to implement our classifier) to extract articles in this period.

which is neutral or positive towards benefit recipients. These results are robust to a variety of specifications and leaving out any one event.

We build a conceptual framework of welfare take-up to explore the mechanisms behind these results. This models the take-up decision being associated with various 'bad' types. There are many different 'bad' types that benefit recipients have been associated with, including low ability, free rider (Friedrichsen et al., 2018), poor (Holford, 2015), and fraudulent (Gavin, 2021). Many of the newspaper articles in our sample contain reference to these types. Stories can affect the probability of being seen as one of these types, the penalty for being seen as one of these types, or the individual's perception of these parameters. Social-image costs are modelled with these parameters with a social reference group, and self-image is modelled analogously with the self as a reference group, following the literature on self-signalling (Bodner and Prelec, 2002; Mijovic-Prelec and Prelec, 2010; Bénabou et al., 2018). The take-up decision then depends on these social- and self-image costs, as well as other costs and benefits associated with applying for the benefit.

We argue that the link between negative newspaper coverage and non-take-up is mediated by both the self-image and social-image costs, which we jointly term the 'stigma cost'. Negative coverage strengthens the association between 'bad' types and the take-up decision, as well as the penalty for being associated with one of these types. We hypothesise that much of the effect is driven by the increased salience of these underlying parameters in the short run, rather than fluctuations in the parameters themselves.

We rule out alternative mechanisms, that might affect the other (non-stigma) costs or benefits in the take-up decision. Non-take-up of social benefits has commonly been attributed to lack of information about eligibility (Bhargava and Manoli, 2015; Anders and Rafkin, 2022) and the complexity costs of applying (Finkelstein and Notowidigdo, 2019). The effect we observe could not be driven by the provision of information about benefits, as this would lead to an increase in applications rather than the decrease that we observe. Stories could also provide information on the application process and change potential applicants' perceptions of the complexity costs of application. This would push the effect in the direction that we observe, however, we find

that none of the articles in our sample discuss the application process, the administration of Universal Credit, or any issues with applications, making this an unlikely mechanism. Many of the articles do discuss cases of benefit fraud. We demonstrate that deterrence of benefit fraud could not account for the magnitude of our results.

We find that the application rate returns to pre-treatment levels after five days, with no detectable increase in applications to compensate for those that were not made directly after a negative story. Therefore it seems that a story can be associated with four days of treatment, on average, with individuals on the fifth day not being treated. This interpretation further supports the saliency mechanism described above. Plausibly, after a four-day period, the increased saliency of the penalty dissipates, as other news stories or topics become more salient. This does not preclude a delay in take-up in the more medium-run of a couple of weeks. However, this is not unusual with any determinant of non-take-up. Much of non-take-up comprises delaying rather than never claiming.

As applications for social benefits are associated with a direct monetary receipt, this allows us to put a direct cost on willingness to pay to avoid stigmatisation, for compliers. The minimum monthly Universal Credit payment in 2021 was £368 (€429, $462) for a single individual, which is approximately £12 (€14, $15) per day.[4] If we conservatively assume that individuals simply delay their applications to Universal Credit by three days then a back-of-the-envelope calculation gives a loss of £36 (€42, $45) per individual, in response to a single story. This is a lower bound as we see no compensating increase in applications after three days, implying that individuals delay for longer, or simply never apply. If a story fully deters an individual from applying, then the cost would go up to £810 (€944, $1,017), the average amount of Universal Credit paid out in 2021.[5]

This paper contributes to several literatures. First, our research sheds new light on the drivers of benefit non-take-up. There is an abundance of qualitative evidence that attests to the existence of welfare stigma. Baumberg (2016) finds that 20.4% of people in the UK think that people should feel ashamed to claim at least one benefit, and 16.9% stated that their take-up behaviour

[4]See https://www.gov.uk/universal-credit/what-youll-get

[5]See https://www.gov.uk/government/statistics/universal-credit-statistics-29-april-2013-to-1
universal-credit-statistics-29-april-2013-to-14-july-2022

would be affected by personal shame.[6] However, survey evidence on stigma is difficult to interpret causally, as the admission of stigma is itself stigmatising, and when multiple factors affect a decision, it is difficult for individuals to report the importance of any specific one of them. Celhay, Meyer and Mittag (2021) find that survey reporting of benefit receipt is particularly flawed. Friedrichsen, König and Schmacker (2018) address these issues by studying the effect of stigma in a lab setting, finding a significant effect on hypothetical benefit take-up. We provide, to the best of our knowledge, the first real-world causal evidence of the effect of stigma on benefit take-up.

Further, we build on the wider literature on non-take-up, which has found that lack of information (Holford, 2015; Bhargava and Manoli, 2015; Anders and Rafkin, 2022) and complexity of application procedures (Finkelstein and Notowidigdo, 2019) are important factors. Another part of the literature considers the role of procrastination in non-take-up, particularly in the context of education subsidies (Dynarski, 2007; Sunstein, 2013; Narayan, 2020). However, while these papers acknowledge the relevance of stigma, they face challenges in isolating it from other drivers of non-take-up. For example, Holford (2015) find peer effects on the take-up of free school meals, but conclude this is primarily an information effect.[7] In our empirical setting, any information effects would act in the opposite direction to the effect of stigma, allowing us to compellingly isolate this effect.

Second, our work adds to the quickly-growing literature on social-image and self-image effects on economic behaviour (see Bursztyn and Jensen (2017) for a review), as well as the large literature in sociology and social psychology which studies and conceptualises stigma (eg. Link and Phelan, 2001; Major and O'Brien, 2005; Pescosolido and Martin, 2015). A clutch of randomised controlled trials study the effect of stigma on outcomes as diverse as voter turnout in the US (Dellavigna et al., 2017), saving behaviour of sex workers in India (Ghosal et al., 2022), and effort in high school (Bursztyn, Egorov and Jensen, 2019a). Perhaps the most similar paper,

[6]Similar surveys have been conducted in other contexts. In the US, Stuber and Kronebusch (2004) report that more than half of individuals surveyed agreed with the statement, "Many people on welfare do not want other people to know they are on welfare" and between 35% and 45% believe that "People in this country on welfare are lazy".

[7]Similarly, Bhargava and Manoli (2015) run a small side experiment on stigma and find negative effects on take-up, but they conclude that this is probably due to a change in perceptions of complexity.

Osman and Speer (2023), looks at participation in job training and a job fair in Egypt, after interventions aimed at alleviating stigma concerns. They find that their interventions actually decrease participation, which they interpret as evidence of a stigma effect. To the best of our knowledge, we are the first to show an effect of any type of stigma in a natural experiment. We additionally contribute to the literature on self and social image by having a natural way of 'pricing' the cost of these concerns, as non-take-up translates directly into forgoing a specific monetary amount.

Finally, we contribute to the literature on media persuasion. A number of papers consider how media can shape attitudes towards marginalised groups. Djourelova (2023) recently looked at the effect on attitudes towards immigrants in the US after the Associated Press banned the use of the term "illegal immigrant", finding significant changes in support for immigration policies. Ivandic, Kirchmaier and Machin (2019) show how newspaper coverage increases the number of Islamophobic hate crimes in the UK after terrorist attacks.[8] DellaVigna, Enikolopov, Mironova, Petrova and Zhuravskaya (2014) provide evidence on radio fostering nationalistic attitudes across the Serbo-Croatian border. While these papers focus on the response of the majority group, we focus on the response of the marginalised, or potentially marginalised, individuals. In doing so, we complement a large sociological literature which has discussed the role of media in the stigmatisation of poverty and welfare (see for instance Morrison, 2019).

More broadly, the literature on media persuasion has focused on the effects of exposure to a particular media outlet or medium (DellaVigna and Kaplan, 2007; Enikolopov et al., 2011; Levy, 2021). Notably Foos and Bischof (2021) also focuses on exposure to *The Sun*, looking at effects on Euroscepticism. In this paper, we open this black box, focusing on the micro-level effects of exposure to individual stories.

The rest of the paper is organised as follows. The following section presents the context of media and benefits in England, and the subsequent one describes our data. Following this, we present evidence on the link between negative media coverage of benefit recipients and take-up of benefits. After this, section 5 describes our conceptualisation of stigma and discusses our

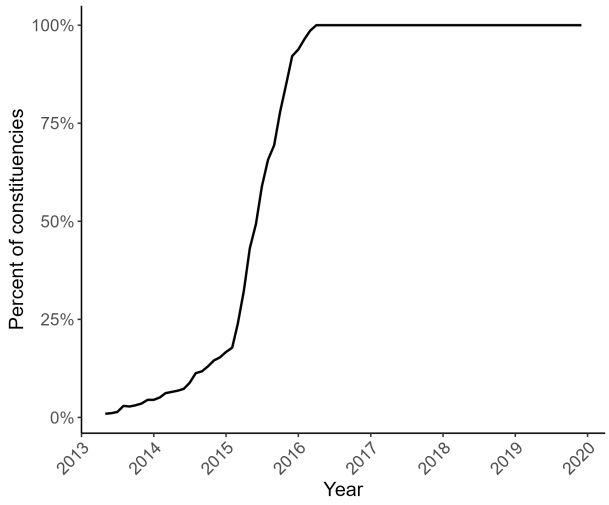
[8] Bursztyn, Egorov, Enikolopov and Petrova (2019b); Müller and Schwarz (2021, 2023) also find that social media can mediate xenophobic opinions and hate crime.

results in light of this, and section 6 concludes.

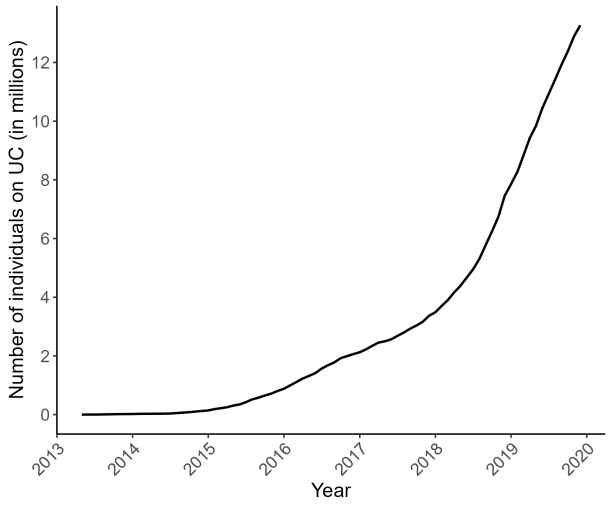
2 Context

2.1 Universal Credit

Universal Credit in England provides an ideal context in which to study non-take-up, as it is a "catch-all" benefit which replaced all working-age means-tested benefits. This allows us to study benefits in general, rather than requiring us to focus on a specific benefit.



(a) Share of parliamentary constituencies that have rolled out Universal Credit live service.



(b) Number of people receiving Universal Credit.

Figure 2.1 – Time series of Universal Credit across constituencies and individuals.

Notes: Sourced from the House Of Commons Library at `https://commonslibrary.parliament.uk/constituency-data-universal-credit-roll-out/`.

Universal Credit was introduced with the Welfare Reform Act of 2012. It is managed by the Department for Work and Pensions. It was initially launched in areas of north-west England starting in April 2013, and then was rolled out progressively across the country. To start with, Universal Credit was only offered to a limited range of claimants, principally single, working-age people, with no children, who were seeking work. This was known as 'live service'. Figure 2.1b shows the roll-out across parliamentary constituencies over the period 2013-2020. In November 2014, 'full service' Universal Credit began to be rolled out, available to all working-

age individuals. By May 2016, live service was available across England[9] and full service was available to all by December 2018.

Universal Credit replaced six means-tested benefits, known as 'legacy benefits'.

1. Jobseeker's Allowance: means-tested unemployment benefit, not related to National Insurance contributions.[10]
2. Income Support: means-tested benefit for individuals with a low income, who are not receiving Jobseeker's Allowance.
3. Housing Benefit: means-tested support for housing costs.
4. Employment and Support Allowance: support for people who have limited capability for work because of their sickness or disability but do not get Statutory Sick Pay.
5. Working Tax Credit: tax credit on low earned income.
6. Child Tax Credit: means-tested tax credit for people with children.

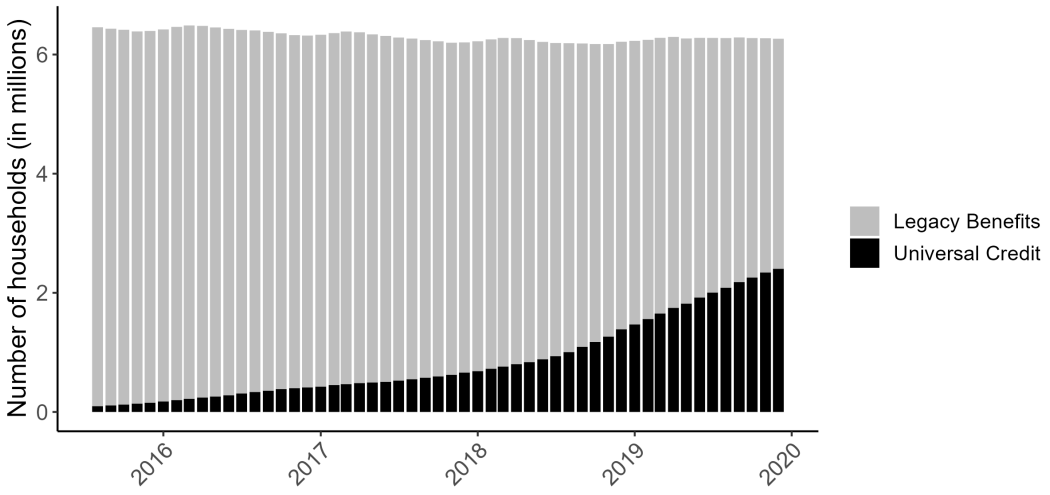


Figure 2.2 – Number of households receiving Universal Credit or legacy benefits.

Notes: Data from the House Of Commons Library at `https://commonslibrary.parliament.uk/constituency-data-universal-credit-roll-out/`.

[9]Our analysis primarily focuses on England, as Scotland and Northern Ireland receive different newspaper editions and data for heterogeneity analysis is not available for Wales. Scotland and Wales rolled out Universal Credit on the same schedule as England, while Northern Ireland, rolled out Universal Credit according to a different schedule

[10]Only income-based Jobseeker's Allowance has been merged into Universal Credit. Contribution-based Jobseeker's Allowance has remained as a separate benefit. Unless specified, when referring to Jobseeker's Allowance, we are referring to income-based.

During both the 'full' and 'live' service phases, applicants could only claim Universal Credit if either they were not at the time receiving any of the legacy benefits or they were receiving a legacy benefit but reported a change in circumstance that changed their eligibility. Claimants have one month after a change in circumstance to report it to the Department of Work and Pensions. Otherwise they are considered as claiming fraudulently.

People who had no changes in circumstance continue to receive legacy benefits, until the final phase of the rollout, 'managed migration', which started in July 2022 and is still underway. Hence, during our period of study, all Universal Credit claimants are either new applicants, or those receiving legacy benefits who had a change of circumstance. The overall number of people receiving either Universal Credit or legacy benefits remained stable across this period (see Figure 2.2).

Eligibility criteria for Universal Credit have remained relatively unchanged, compared to the legacy benefits that it replaced, meaning that individuals receiving Universal Credit can be categorised according to which types of legacy benefit they would have received.

Figure 2.3 shows the number of households (or individuals) that receive Universal Credit or legacy benefits by type of benefit. As of May 2021, 62% of households receiving working-age, means-tested benefits were receiving Universal Credit, with notable differences across benefit types. 96% of individuals receiving unemployment benefits were receiving Universal Credit, whereas only 34% of those receiving disability-related benefits were receiving Universal Credit. This suggests that the turnover of individuals on unemployment benefits is far higher than that of disability benefits. Back-of-the-envelope calculations suggest that there are at least four times as many new recipients of unemployment benefits, on a given day, as there are for disability benefits, despite similar numbers of recipients overall.

The Department for Work and Pension have not published take-up statistics of Universal Credit. However, they have continued to publish data on take-up of legacy benefits, which is shown in Figure 2.4. Take-up of Universal Credit may be somewhat higher, as one of the key motivations for the switch to Universal Credit was that a simpler benefit system would increase take-up.[11]

[11] `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/`

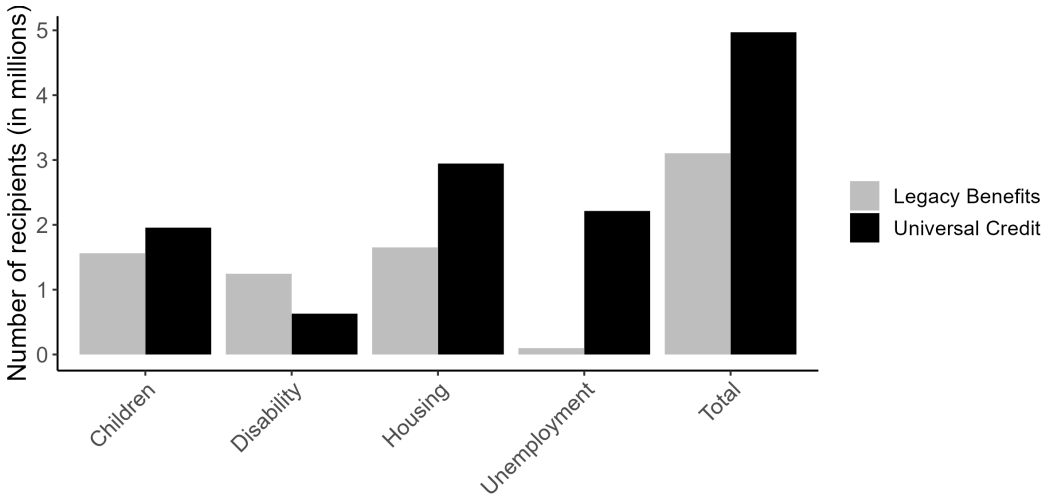


Figure 2.3 – Recipients of Universal Credit or legacy benefits, by type of support received.

Notes: Unemployment benefits are measured as number of individuals. All others are measured as number of households in May 2021. Sourced from the House Of Commons Library at `https://commonslibrary.parliament.uk/constituency-data-universal-credit-roll-out/`.

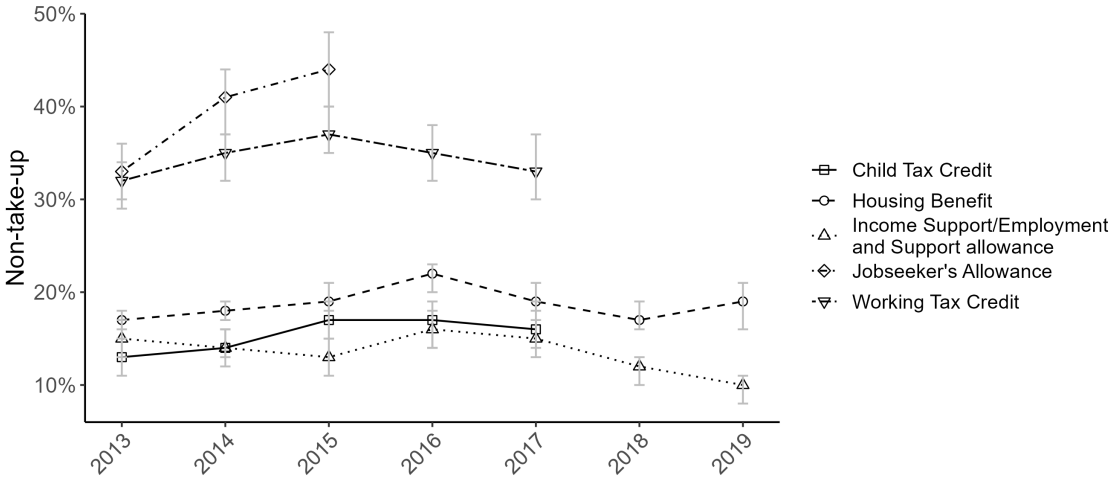


Figure 2.4 – Non-take-up of legacy benefits.

Notes: Sourced from the Department of Work and Pensions at `https://www.gov.uk/government/collections/income-related-benefits-estimates-of-take-up--2`. The numbers presented are based on caseload: "Caseload take-up compares the number of benefit recipients, averaged over the year, with the estimated number who would be receiving if everyone took up their entitlement for the full period of their entitlement."

Figure 2.5 – Selection of *Sun* front-pages with negative content on benefit claimants.

2.2 English Newspapers

The newspaper landscape in England is commonly divided into two categories: 'broadsheets' (eg. *The Telegraph*, *The Times*, *The Guardian*) and 'tabloids' (eg. *The Sun*, *The Daily Mail*). Historically this was a distinct divide based on the size of the paper. As many broadsheets have more recently opted for a more compact size, this has become a shorthand for the type of coverage; broadsheets are typically seen to contain more 'quality' content, while tabloids are renowned for their sensationalist coverage. An emerging third category are freesheet newspapers (eg. *The Metro*, *The Evening Standard*), which are distributed in urban centres without charge. The reporting style of tabloid newspapers provides an excellent context for studying the persuasive effects of media, as opinionated content is not restricted to editorials and columns; value-laden

attachment_data/file/48897/universal-credit-full-document.pdf

language is often used in front-page news. Figure 2.5 shows various examples.

Further, the circulation of tabloid newspapers is nearly three times that of broadsheets (figure 2.6), demonstrating the salience of the tabloid press and its potential to create and enforce social norms.

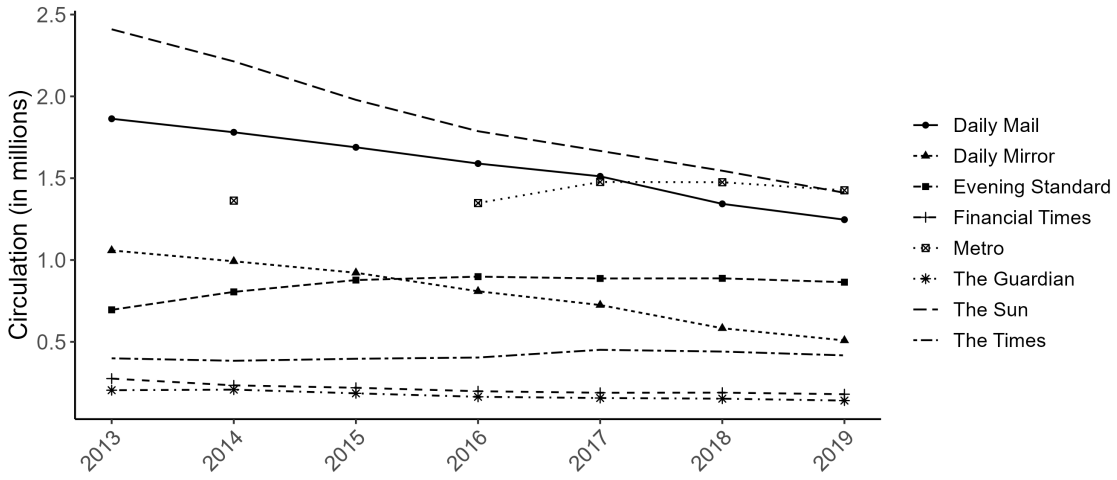


Figure 2.6 – Circulation of major newspapers in the UK

Notes: Sourced from the Audit Bureau of Circulations, reported in the Press Gazette.

By circulation, the largest of the tabloid newspapers is *The Sun*. *The Sun* is a daily tabloid, famous for topless page-three models and sensationalist headlines of questionable veracity. The paper's parent company, News UK, is owned by Rupert Murdoch's NewsCorp. A standard edition cost £0.40 (€0.47, $0.63) in 2013 and £0.65 in 2019 (€0.75, $0.90). As separate regional editions, with different content, are published for Scotland and Northern Ireland,[12] we focus our analysis on the National Edition, which covers England and Wales.

The Sun has made bold claims to influence British opinion, most notably claiming responsibility for the 1992 Conservative general election victory, with the headline "It's The Sun Wot Won It". Using sentiment dictionaries, a 2012 study by anti-poverty think tank, Turn2Us, found that out of all British newspapers, *The Sun* ran the highest volume of articles containing negative content concerning benefit recipients.[13] In this paper we study whether this negative content influences the opinions and actions of the British public.

[12] *The Sun* also has a Republic of Ireland edition.

[13] https://www.theguardian.com/news/datablog/2012/nov/20/benefits-stigma-newspapers-report-we

3 Data

Two main sources of data were used for this study: data on daily applications to Universal Credit and data on articles that contain negative content about benefit recipients from *The Sun*. This section describes these datasets and provides some descriptive statistics.

3.1 Data on Universal Credit applications

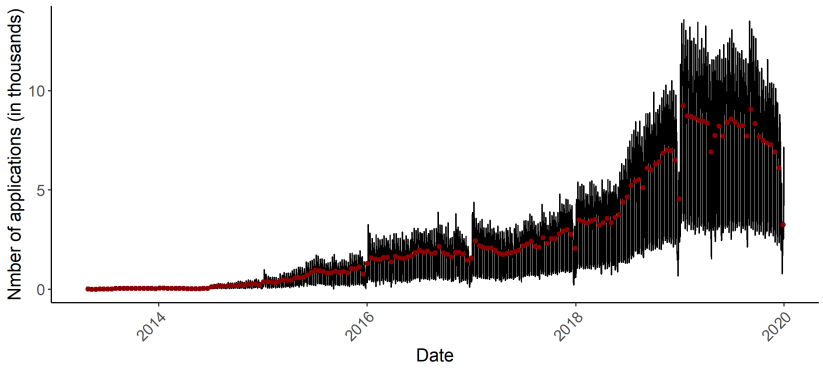
We use administrative data on the daily number of applications to Universal Credit over the period 2013-2019. This data is broken down at the postcode area level (first two postcode characters). There are 106 postcode areas. Most of the major cities have their own postcode area (except London which is separated into 8) with populations ranging from around 30,000 to 2,000,000 per area. We sourced the data from the Department for Work and Pensions.[14] It records the number of applications made online, on a given day, regardless of whether they were successful or not. Universal Credit applications can only be made online.

In figure 2.7a, we show daily applications over the period. There is an upwards trend, which is due to the progressive roll-out of Universal Credit, described in section 3. The variance of the data is caused by a regularly weekly pattern in applications. In figure 2.7b, we zoom into a 3-week period in February and March 2016 to illustrate the weekly trend. The number of daily applications decreases more or less monotonically across the week, with the highest numbers on Monday, and then substantially lower numbers during weekends, as shown in 2.7c. In addition, we see lower numbers of applications at the end of each year, during the Christmas and New Year holiday period.

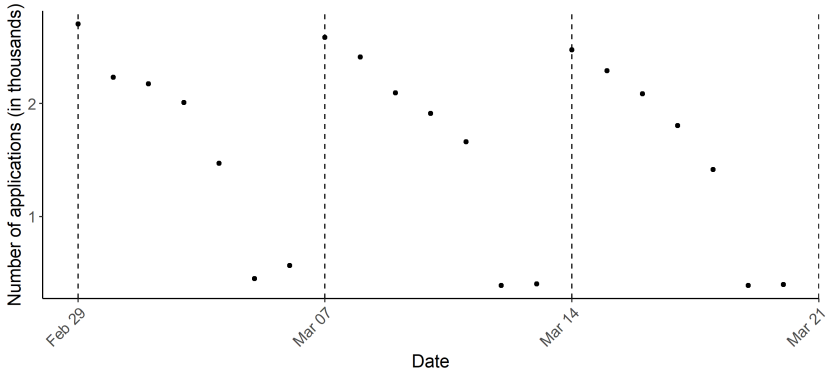
3.2 Data on newspaper articles

Between 2013 and 2019, *The Sun* published 1.7 million articles, making it unfeasible to categorise which articles contain negative content about benefit recipients by hand. We developed two alternatives for extracting this subset of articles. First, we use a keyword method, classifying

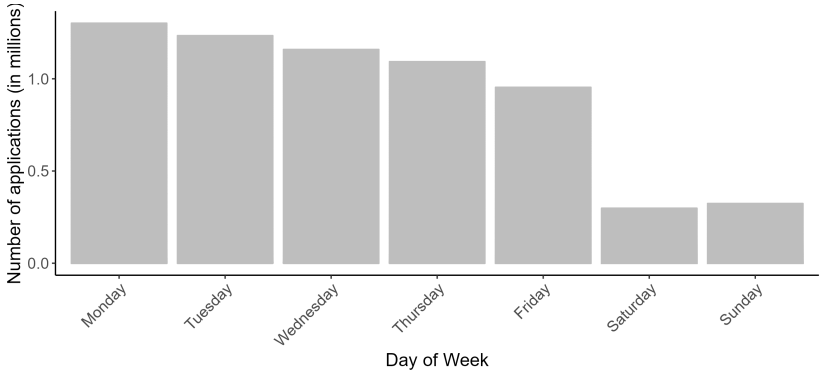
[14]The data was previously available at `https://stat-xplore.dwp.gov.uk`, but has since been removed.



(a) Daily Universal Credit applications, 2013-2019. Red dots show weekly averages.



(b) Daily Universal Credit application over the 3 weeks between February 29th and March 20th 2016. The dashed lines are Mondays.



(c) Total number of Universal Credit applications by day of week, 2013-2019.

Figure 2.7 – Daily Universal Credit applications

Notes: Panel 2.7a shows the overall trend of the log of the number of applications. Panel 2.7b shows a randomly chosen three-week interval between February 29th and March 20th 2016. Panel 2.7c shows the day-of-week pattern of Universal Credit applications.

an article as containing negative content about benefit recipients if it contains one of a set of keywords. We use keywords motivated by previous work in sociology (see Morrison (2019)

and Appendix 3 of Geiger et al. (2012)), namely “benefit fraud", “scrounger", “shirker", “skiver", “sponger", “sponging", “welfare cheat" and “workshy".

Second, we leverage a contextualised language model and finetune a classification head to sort articles. We labelled 400 articles (173 positive and 227 negative) about whether they were mainly on the topic of social benefits. We used this data to finetune distil-RoBERTa. RoBERTa (Liu et al., 2019) is an updated BERT-type Large Language Model, trained on ten times as much training data as BERT (Devlin et al., 2019), with an improved training methodology. Distil-RoBERTa (Sanh et al., 2020) distils RoBERTa from 125 million parameters to 82 million, with minimal loss in performance. [15] We run this over concatenated headlines, subheadings and abstracts.[16]

Under a tenth of one percent of the articles published during this period are about social benefits, meaning it is feasible to manually check all the articles that are classified as ‘on topic’, so we can totally eliminate false positives (type I errors) from either classification method. Therefore when selecting the better method, we are primarily concerned with false negatives (type II errors); we do not want to classify articles as ‘not on topic’ when they should have been classified as ‘on topic’, but we are not too worried about classifying articles as ‘on topic’ when they should be classified as ‘not on topic’, as these errors can easily be manually fixed. Missing ‘on topic’ articles can be a particular concern for our identification strategy (see section 4.1) as they could lead us to include days that were actually treated in our pre-period.

In practice, we found that the finetuned distil-RoBERTa produced far better results than using keyword searches. The keyword we used were chosen to detect articles with negative content, rather than all articles about benefit seekers, so the relevant point of comparison is the number of articles retrieved containing negative content, rather than the number retrieved that were mainly about benefit seekers. Over 2014-2016, the language model found 246 articles that contain negative content about benefit recipients. Keyword searches only found 110 (44.7%) of these. Further, the articles found by keyword search were a strict subset of those found by distil-RoBERTa. Additionally, we found that the keywords used in the literature were particularly

[15] The small size of Distil-RoBERTa makes it feasible to run inference on the single CPU provided by TDM Studio.
[16] Separated by RoBERTa’s [SEP] token.

likely to be used in commentary and editorials. These are often published the day after a news article containing negative content about benefit seekers, but often this news article is missed by the keyword classification method. In these cases, we would end up counting the first day of the story (ie. the first treated day) as the day of the editorial, and counting the previous day as part of the (untreated) pre-period. In the cases where there was actually an article on that day, this leads to contamination of the pre-period. Thus we prefer the language model classification where possible. Table 2.1 shows examples of articles that are picked up by the language model, but not by keyword searches.

Table 2.1 – Text classification by a keyword search and a language model.

Article	Keywords	Language model
Living it Lard on Benefits: 'Fattie' in £100k swindle (2013-06-04) A FORMER fattie kept claiming almost £100,000 in handouts — after shedding 19 stone. [...] Shellard began losing weight in 2008. But she failed to mention her improving health so that she could continue receiving disability living allowance, income support, housing benefit and council tax benefit totalling £98,000. [...] Judge Mr Recorder Sephton QC told her at Manchester Crown Court: "Many people would regard you quite rightly as a scrounger — and a lying scrounger at that."	On topic	On topic
Benefits cap calls (2016-07-09) A PETITION demanding a benefits cut for "Britain's most shameless mum" has topped 15,000 signatures. Mum-of-12 Cheryl Prudham, 33, reportedly rakes in £40,800 in handouts. But the part-time cleaner, of Lancashire, said on telly she is saving for a boob job and wants baby 13 with a sperm donor after dumping her hubby. It prompted critics to set up a change.org petition to get her cash cut.	Not on topic	On topic

Notes: The first example has been abridged.

We collect articles from *The Sun* using two sources. We use ProQuest for articles between 2013 and 2016 and LexisNexis for 2017 to 2019. In both cases, we restrict to articles from *The Sun*'s National Edition, which covers England and Wales. Up until the end of 2016, ProQuest contains data on all *Sun* articles, including headlines, subheading, editions, page number, and importantly a short extractive summary of the article. They do not have full text data from the articles. By contrast, LexisNexis contains full text data on all articles during our period of study. ProQuest data has the advantage of allowing direct import into TDM Studio, a software which

enables us to run inference from our finetuned language model. By contrast LexisNexis, which contains full text data from the whole period, only allows extraction of articles via keyword search.

Therefore for 2014-2016, we extract articles using our finetuned language model. From 2017 onwards, ProQuest stopped processing extractive summaries. We found that headlines and subheadings alone were not enough to reliably extract articles about social benefits. Therefore for 2017-2019, we rely on keyword searches over the full text data in LexisNexis.

For both sets of data, we then manually reviewed all extracted articles, and discarded those that were not mainly about social benefits. We further manually classify each article according to a number of dimensions. First and most importantly, we identify whether an article is explicitly negative about benefit seekers. We do not count articles with implicit negative content, although we recognise that these might also affect take-up behaviour. We also do not count articles that contain information about policy changes that might be interpreted as negative towards benefits. We further remove articles that were negative about benefit seekers, but focused on two 'out groups': immigrants and Islamic extremists. In the former case, we saw a number of articles related to the possibility of Eastern European immigrants coming to the UK to seek benefits. In the latter, there were a number of articles, mainly related to one individual, calling for the removal of benefits for individuals who had pledged allegiance to the Islamic State. In both cases, the articles focused on groups that *Sun* readers would generally not feel themselves to be members of, making the expected effect of articles unclear. Readers may still perceive negative sentiment towards all benefit seekers, or they may interpret such articles as meaning benefits should belong to the 'in group', making them more likely to apply. We also do not include reader letters or articles related to television programmes.[17] Appendix A.1 presents some representative articles from our dataset.

Overall, we retrieve 1,809 articles from our two classification methods (1,679 from language model classification in 2014-2016 and 130 from keyword classification in 2017-2019). Manual

[17] In particular, in 2014, Channel 4 released a controversial documentary called 'Benefits Street', which focused on the lives of several families on benefits. *The Sun* ran multiple stories about these individuals, both related to their benefit-related behaviour and their wider lives.

Table 2.2 – Descriptive statistics by year on the number of articles.

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**Total**
All articles (in thousands)							
	317	288	297	278	276	256	**1,712**
Articles predicted as about benefits							
LLM	683	564	432	-	-	-	**1679**
Keywords	(158)	(159)	(73)	55	43	32	**520**
Articles manually classified as about benefits							
LLM	287	219	121	-	-	-	**627**
Keywords	(50)	(71)	(24)	9	12	6	**172**
Removal of articles about TV, policy and outgroup							
LLM	164	140	68	-	-	-	**372**
Keywords	(32)	(57)	(24)	9	12	5	**139**
Articles manually classified as negative							
LLM	102	97	47	-	-	-	**246**
Keywords	(30)	(55)	(24)	9	10	5	**133**
Unique days with a negative article							
LLM	75	68	40	-	-	-	**183**
Keywords	(22)	(30)	(17)	9	10	4	**92**
Unique days after overlap restriction							
LLM	6	11	11	-	-	-	**28**
Keywords	-	-	-	6	9	3	**18**

Notes: This table shows the number of newspaper articles for each year and in total after every restriction. The first line gives the total number of Sun articles, rounded in thousands. The second line gives the number of articles that our two methods, keyword search and the Large Language Model as described in the text, extract. The next line shows the numbers after removing articles with TV, policy or outgroup related mentions of benefits. The two following lines show the number of articles that are actually about benefits and then that are negative towards benefit seekers. The penultimate line shows the number of days with a negative article (ie days with several negative articles are counted only once). The last line shows the number of articles after restrictin to 10-day windows with one-sided overlap as described in section 4.1 for the main regression.

verification of these found that 654 of these were in fact mainly on the topic of social benefits. 270 of these were manually classified as explicitly negative, across 206 distinct days (an average of 1.3 articles per day).

Table 2.2 shows these figures broken down by year and classification method. The number of negative articles peaks in 2014, and then decreases thereafter, which matches the patterns

described elsewhere (Morrison, 2019). During the period where we have a reliable measure of the total number of articles about social benefits, negative articles make up 66% of these, with the vast majority of the rest being neutral. Only 0.8% of articles were labelled as explicitly positive.

3.3 Other data

We also use information on income by region for our heterogeneity analysis. We obtain data on income at the constituency level from official government sources.[18] Constituencies are the geographical entities that elect a single Member of Parliament. There are 650 constituencies in total, with 533 constituencies in England, 40 in Wales, 59 in Scotland and 18 in Northern Ireland (as of 2019). The number of eligible voters varies roughly between 50,000 and 100,000 per constituency. Constituencies and postcode area (regional decomposition in the Universal Credit data) are not uniquely matched.[19] For each postcode area, we therefore take a weighted mean of income of the constituent constituencies, weighted by number of individuals.

4 The effect of stories on benefit take up

In this section, we examine the effect of negative news stories on applications for Universal Credit. We start by describing our empirical approach and then discuss our results as well as informative heterogeneity.

4.1 Event Study

We estimate the effect of negative stories on the number of applications for Universal Credit using an event study in the window around the date of publication of the first article in a negative story.

Specifically, we estimate regressions of the form:

[18] See `https://www.gov.uk/government/statistics/income-and-tax-by-parliamentary-constituency-2`
[19] 41% of constituencies have a unique postcode area, 32% have two and 16% have 3.

$$\ln\left(\text{Applications}_{e,\tau}\right) = \alpha + \sum_{\tau \in \{-\underline{w},\ldots,\overline{w}\}/\{-1\}} \beta_\tau D_\tau + \nu_{e,\tau}$$

where e represents an event which is defined as a window of days centered at the date of publication of a negative story with $\underline{w}$ days before and $\overline{w}$ days after. Relative time within the window is denoted with τ. The variable D_τ corresponds to the relative time dummy. The reference day is the day before publication of the first article in a story.

$\text{Applications}_{e,\tau}$ is the total number of applications made to Universal Credit on day τ of event e. We only include applications made in England. We exclude Scotland and Northern Ireland as they receive different editions of *The Sun* and we exclude Wales for consistency with the subsequent heterogeneity analysis for which we do not have data on Wales. We choose a logarithmic functional form to obtain relative effects.[20] This is particularly sensible in our context with a strong time trend in the underlying Universal Credit applications data, meaning that levels change significantly over our period.

We cluster standard errors by event.

Choice of window We define the event day as the date on which a negative article appears in *The Sun*. An event window is defined as $\underline{w}$ days before (the 'pre-period') and $\overline{w}$ days after this event day (the 'post-period').

These windows frequently overlap, when one event falls with the pre- or post-period of another event. As discussed in section 3.2, this is frequently multiple articles that related to the same story, with a lead article on the story on one day, followed by commentary on the same story on the following day.

These overlaps threaten our identification of the effect of a given article when another article, or the post-period of another article, appears in its pre-period, as the supposedly untreated days are actually treated. Therefore we do not consider any articles whose pre-period overlaps with the event window of another article (Morales, 2021). This is not a concern for articles whose post-periods overlap with the event window of another article, as the days in that post-period

[20] The result in levels is discussed in section 4.4.

remain treated. Thus, in our preferred specification, we do not remove an article if the event window of another article overlaps with its post-period. This is illustrated in figure 2.8. Further, due to the pattern of clusters of articles about the same story, dropping such articles would dramatically reduce our sample size. It would also select the smallest - and least salient stories - as it would select those that receive no follow-up article. The 'one-sided' window selection naturally selects the first article of a story for the treatment date. Thus our estimates can be properly interpreted as the effect of a story, or article cluster, rather than an individual article.

Nonetheless, our results are robust to a 'two-sided' window selection, where articles are only used if their event period does not overlap with that of any other article in the pre- or post-period. Results of this are shown in figure 2.23.

(a) Event windows included in analysis in 'one-sided' case. In this case the first article in a story cluster is kept.

(b) Event windows included in analysis in 'two-sided' case. In this case, all articles within a story cluster that contains more than one article are lost.

Figure 2.8 – Illustration of different ways of accounting for overlapping event windows. Green boxes represent those event windows that are retained in the analysis and red boxes represent those that are removed.

A second choice is the length of the pre-and post-period. As the length of the window determines which articles are kept and which are dropped, the choice of window length presents a trade off. If the windows are too small, then we will not remove events that actually had another event whose effects are still affecting the pre-period. If the windows are too large, then we remove unnecessary events, reducing power.

For our main results, we use a pre-period of five days before and a post-period of four days, giving a total of five untreated and five treated days (including the event day), which balances between these two arguments. We find that the effect of a story can be seen for three to four days, so this period is long enough to avoid pre-period contamination, while also avoiding the unnecessary removal of too many events. This gives 46 events in our preferred specification, which is also shown in table 2.2. Robustness to this choice is considered in figure 2.22.

Identifying assumption This empirical strategy exploits the high frequency nature of our data on Universal Credit applications. Our identifying assumption is that on average, the number of application on the days following the first article in a story would have been the same as those on the days prior to the article, if the article had not been published. This assumption is valid if the publication of negative articles is locally random (although this is not a necessary condition). That is, although the publication of negative articles about benefit seekers is unlikely to be exogenous to long-term trends in the number of applications, all that we require is that they are exogenous to changes in the number of applications in the short window around the publication of the article.

There are three threats to this identification strategy. First, we may be concerned about reverse causality. However, we do not find it plausible that newspapers publish negative articles about benefit seekers because they anticipate that there will be a short run fluctuation in the number of applications on a given day, relative to the previous ones.

Second, there may be measurement error concerns. It is possible that some of the articles appear the day before on the *Sun* website. This effect is likely small as articles appearing during the day are likely to have little influence on application behaviour. Additionally, the bias it would introduce would decrease the measured effect and hence make our results a lower bound.

Third, we may be concerned about high-frequency omitted variables that both increase the probability of a negative article being published and decrease the number of applications. For this reason, we exclude from our analysis periods around bank holidays from the analysis, as these affect benefit applications unpredictably and are hard to control for because of their low numbers.

Other omitted variables may include big events such as elections, sport events or natural disasters. These are likely to decrease the amount of news space dedicated to coverage of social benefits. They may also decrease the number of applications for social benefits, if individuals are voting in elections, watching sports events, or recovering from natural disasters. This would lead to a positive bias in results. Given that we find a negative effect of news articles on applications, this would bias our results towards zero, so these types of omitted variable are not a big concern.

Residualisation A further high-frequency variable is the day of the week. As shown in figure 2.7, applications to Universal Credit exhibit strong, but regular, weekly patterns, which may confound our results if negative articles are more likely to be published on particular days of the week.

Additionally, we show in figure 2.9 the day-of-week distribution of negative articles. When considering all the articles, the distribution looks fairly uniform, albeit with a potential increase on the weekend. When considering only articles within a 10-day window, with a one-sided overlap (our main specification), the distribution is noisier, with a peak on Saturdays and Tuesdays.

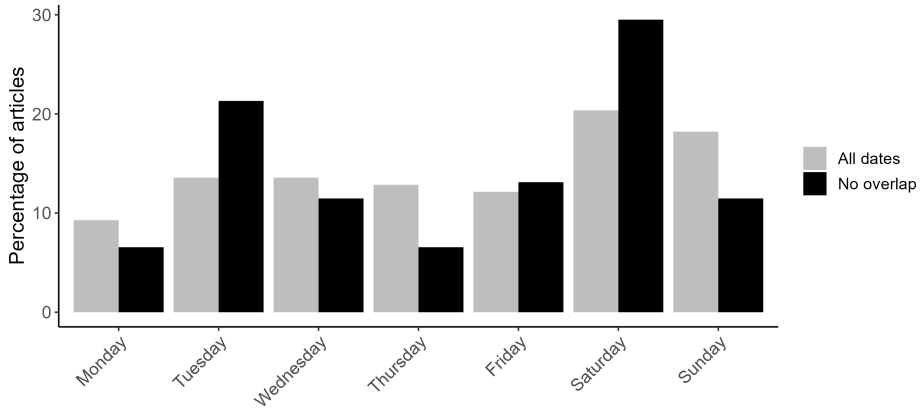


Figure 2.9 – Day-of-week distribution of articles that are negative towards benefit recipients.

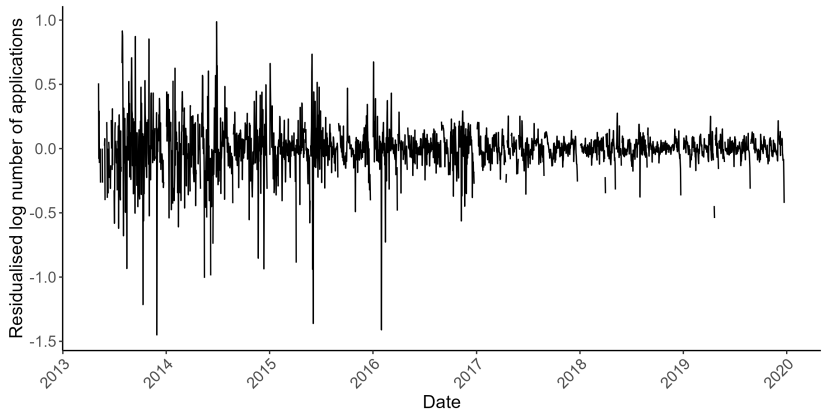
Notes: This figure shows the day-of-week distribution of articles that are negative towards benefit recipients. The grey bars represent the distribution for all such articles. The black bars correspond to the distribution after having restricted the choice of windows around events to limit overlap, as described in section 4.1. We make the same choice as for the main results, namely 10-day windows with one-sided overlap.

To overcome this, we modify our base regression to a two-step procedure: we first residualise the data to eliminate the overall time trend and the day-of-week pattern. We then run the event study on the residualised data.

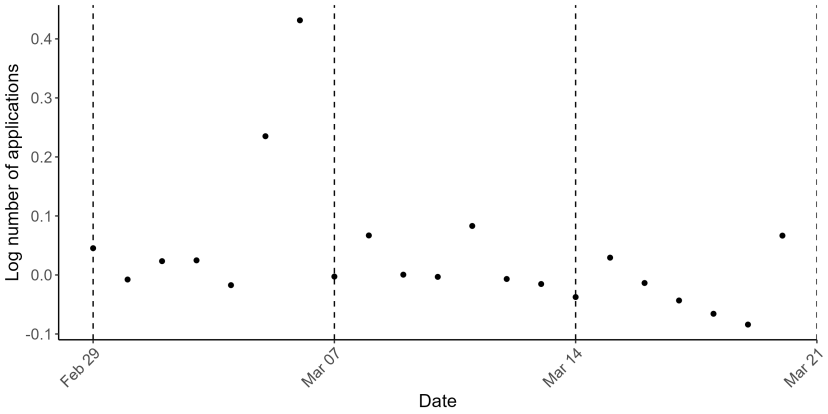
We residualise the logarithm of number of applications using the following specification:

$$\ln\left(\text{Applications}_t\right) = \text{dow}_t \times \text{month}_t \times \text{year}_t + \varepsilon_t \tag{2.1}$$

where dow_t, month_t and year_t are day-of-week, month and year fixed effects respectively. We



(a) Daily Universal Credit applications, 2013-2019, after residualisation.



(b) Daily Universal Credit application over the 3 weeks between February 29th and March 20th 2016, after residualisation. The dashed lines are Mondays.

Figure 2.10 – Daily Universal Credit applications after residualisation.

Notes: Panel 2.10a shows the overall trend of the log of the number of applications after residualising using equation (2.1). Panel 2.10b shows a randomly chosen three-week interval between February 29th and March 20th 2016, after the same residualisation. This can be compared with Panels 2.7a and 2.7b to see the effect of the residualisation.

then compute the residuals from this regression $\overline{\ln(\text{Applications}_t)}$.

We residualise before subsetting the data to the dates in the windows around the publication of articles, which allows us to leverage all dates in the sample and therefore have a much better estimation of the day-of-week pattern, making this two-step approach preferable to adding fixed effects in our event study regression, which could only estimate the day-of-week pattern from a limited number of days, half of which are treated.

In the second step, we estimate the event study as before, except using the residualised

outcome:

$$\overline{\ln\left(\text{Applications}_{e,\tau}\right)} = \alpha + \sum_{\tau \in \{-\underline{w},\ldots,\overline{w}\}/\{-1\}} \beta_\tau D_\tau + \nu_{e,\tau}$$

Other alternatives for residualisation are considered in section 4.4, but yield nearly identical results.

4.2 Results

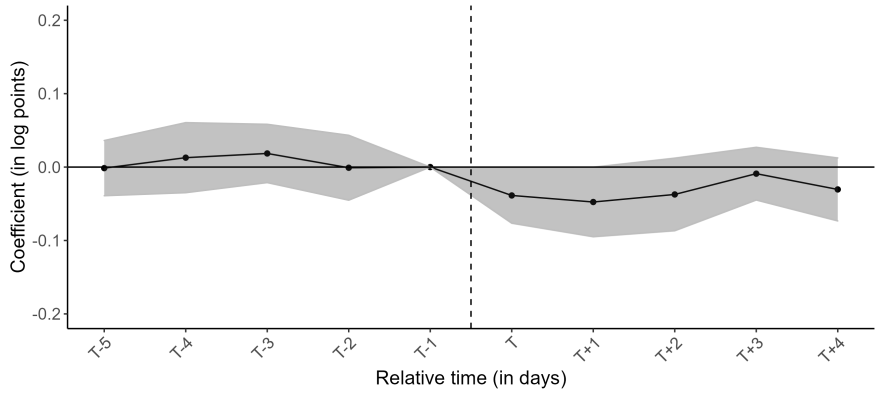
Main result Figure 2.11a shows the result from the event-study specification described in the previous section. A clear pattern is evident in this graph. The pre-trends in the five days preceding the publication of the story with negative content about benefit seekers are flat and statistically insignificant, lending credence to our identification hypothesis. Relative to the day before the publication of the story, there is a drop of 4-5% in the number of Universal Credit applications in the three days following the publication. On the fourth day, the number of applications has returned to normal. This decrease in the first two days is marginally significant at the 5% level with p-values of 0.051 and 0.055 respectively.

The average number of daily Universal Credit applications in England in 2019 was 6,837, which means that over the three-day period a single stigmatising story deters more than a thousand individuals from applying. We discuss the magnitude further and provide back-of-the-envelop computations about the implied social cost in section 5.5.

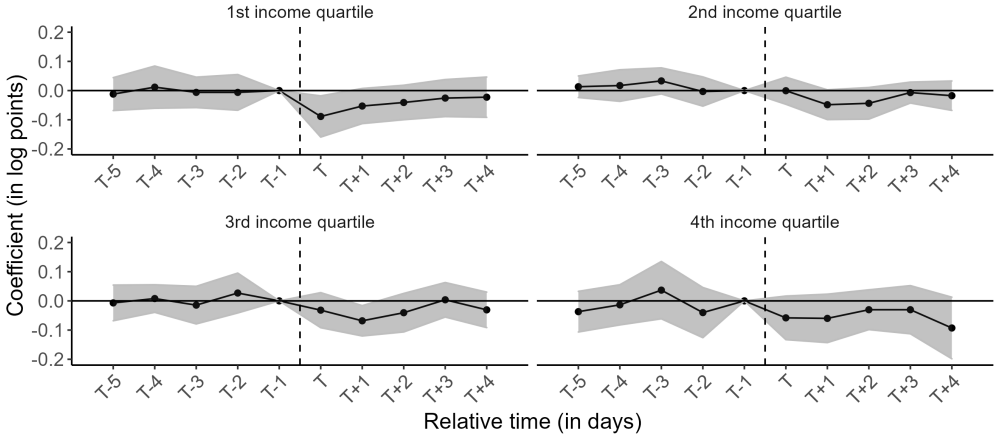
Heterogeneity by regional income level Next, we look at the heterogeneity by average income level in the region. This is interesting for several reasons. First, the rate of eligibility for Universal Credit in low income regions is likely to be higher, which would imply a larger effect in low-income regions.

Second, the majority of *Sun* readership is lower income, which means that inhabitants of lower-income areas are more likely to have been exposed to *The Sun* on the day of the negative story. This can have a direct effect - those thinking about applying for benefits are more likely to be exposed to articles and be deterred from applying - and an indirect effect - people in these areas have a higher expectation that those around them will have seen the articles. However this

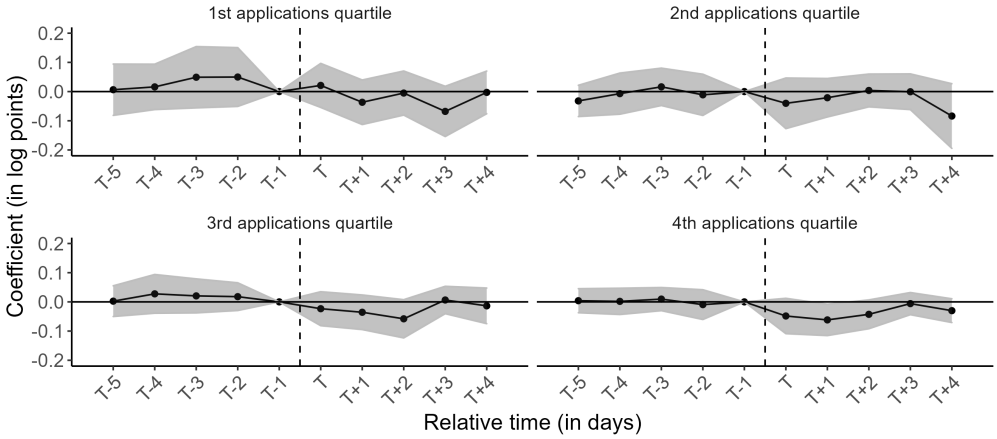
latter effect could also go the other way. Levels of stigma might already be higher in low income regions, meaning the marginal effect of an additional negative story is lower.



(a) Main regression



(b) Heterogeneity by regional quartile of income



(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.11 – Main results from event-study design

Notes: Panel 2.11a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 10-day windows, with only one-sided overlap between events allowed. Panels 2.11b and 2.11c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.

Table 2.3 – Regression table for the main event-study and heterogeneity along income quartiles and application quartiles

	Main regression	Income Quartiles				Application Quartiles			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\tau = -4$	-0.0014	-0.0120	0.0132	-0.0070	-0.0370	0.0064	-0.0320	0.0025	0.0040
	(0.0192)	(0.0287)	(0.0188)	(0.0312)	(0.0355)	(0.0448)	(0.0273)	(0.0268)	(0.0210)
$\tau = -3$	0.0128	0.0117	0.0172	0.0080	-0.0135	0.0160	-0.0071	0.0275	0.0018
	(0.0244)	(0.0369)	(0.0276)	(0.0241)	(0.0351)	(0.0397)	(0.0360)	(0.0338)	(0.0230)
$\tau = -2$	0.0186	-0.0062	0.0331	-0.0146	0.0368	0.0491	0.0163	0.0205	0.0095
	(0.0203)	(0.0266)	(0.0228)	(0.0328)	(0.0501)	(0.0536)	(0.0327)	(0.0297)	(0.0202)
$\tau = -1$	-0.0009	-0.0062	-0.0032	0.0269	-0.0401	0.0498	-0.0112	0.0179	-0.0094
	(0.0226)	(0.0312)	(0.0255)	(0.0348)	(0.0438)	(0.0514)	(0.0360)	(0.0242)	(0.0260)
$\tau = 1$	-0.0386*	-0.0887**	-0.0005	-0.0318	-0.0582	0.0211	-0.0402	-0.0234	-0.0483
	(0.0193)	(0.0360)	(0.0237)	(0.0306)	(0.0381)	(0.0385)	(0.0442)	(0.0297)	(0.0309)
$\tau = 2$	-0.0476*	-0.0530*	-0.0485*	-0.0683**	-0.0600	-0.0366	-0.0211	-0.0356	-0.0617**
	(0.0242)	(0.0307)	(0.0260)	(0.0265)	(0.0423)	(0.0389)	(0.0336)	(0.0302)	(0.0274)
$\tau = 3$	-0.0373	-0.0408	-0.0436	-0.0405	-0.0303	-0.0051	0.0038	-0.0580*	-0.0426*
	(0.0253)	(0.0301)	(0.0276)	(0.0337)	(0.0348)	(0.0385)	(0.0286)	(0.0334)	(0.0252)
$\tau = 4$	-0.0090	-0.0257	-0.0069	0.0037	-0.0303	-0.0679	-0.0006	0.0064	-0.0059
	(0.0183)	(0.0324)	(0.0183)	(0.0302)	(0.0421)	(0.0439)	(0.0312)	(0.0239)	(0.0193)
$\tau = 5$	-0.0305	-0.0227	-0.0175	-0.0307	-0.0929*	-0.0028	-0.0838	-0.0136	-0.0301
	(0.0219)	(0.0352)	(0.0257)	(0.0310)	(0.0538)	(0.0372)	(0.0565)	(0.0311)	(0.0208)
Observations	460	460	450	450	370	350	420	450	460
R^2	0.03077	0.02625	0.02807	0.02273	0.03478	0.03722	0.02036	0.02583	0.02949
Adjusted R^2	0.01138	0.00677	0.00819	0.00274	0.01065	0.01173	-0.00114	0.00590	0.01008
N. of events	46	46	45	45	37	35	42	45	46

Notes: Column shows the result from the two-step procedure described in the text. The data is residualized for day-of-week, month and year fixed effects using the whole data. We then restrict to 10 day windows around the publication of a negative article about benefit claimants. Columns 2 to 5 and 6 to 9 show the heterogeneity by income and Universal Credit application quartiles respectively. All errors are clustered at the vent level.
*p<0.1; **p<0.05; ***p<0.01.

To test this, we classify postcode areas into average income quartiles in 2012. We then apply the two-step procedure described above to each quartile separately.[21] The results from this are shown in figure 2.11b. This figure shows that our results are the strongest in the quartile with the lowest average income. This quartile exhibits completely flat pre-trends followed by an economically and statistically significant drop of 9% in applications on the day of publication of the first article in a story. Levels of applications progressively return towards the pre-period levels, and arrive there on the fifth day. The second and third quartiles exhibit a similar pattern with a smaller magnitude. In the fourth quartile, numbers of applications are quite low, making results quite noisy, and there is no clear effect of negative story.

Heterogeneity by number of applications To confirm this heterogeneity, we also look at heterogeneity according to regional level of applications. The effect of this heterogeneity is ex ante unclear because there could be a density effect which makes claiming more "acceptable", but for instance Baumberg (2016) argues and finds that there is more stigma in high-claim areas.

As before, we split the data into quartiles of average number of Universal Credit applications in 2019 and apply the described two-step procedure. Figure 2.11c shows the results. Similarly to the previous analysis, the effect is entirely driven by the fourth, highest-claim quantile, with small, insignificant effects in quartiles two and three, and a noisy quartile four, with little indication of an effect.

Heterogeneity by article position in newspaper Another possible dimension of heterogeneity is the role of article position in the newspaper. Articles are more likely to be read if they are closer to the front of the newspaper, especially if they are on the front page. It seems likely that the earlier the article appears in the newspaper, the more people are potentially exposed to it and hence treated. Additionally, earlier articles generally tend to be on (what *The Sun* editors consider to be) more sensational stories, which could be expected to generally be more negative towards benefit seekers.

Ideally, we would want to present page number heterogeneity with a restriction to a certain early subset of pages. Given the small number of events and the fact that the chosen articles are fairly evenly distributed across pages (see appendix figure 2.17), we choose instead to present

[21] In particular, the residualisation is now done for each quartile separately. Additionally, given the progressive roll-out of Universal Credit, the quartiles do not have the same number of events because some events happen while Universal Credit had not been rolled out to those regions

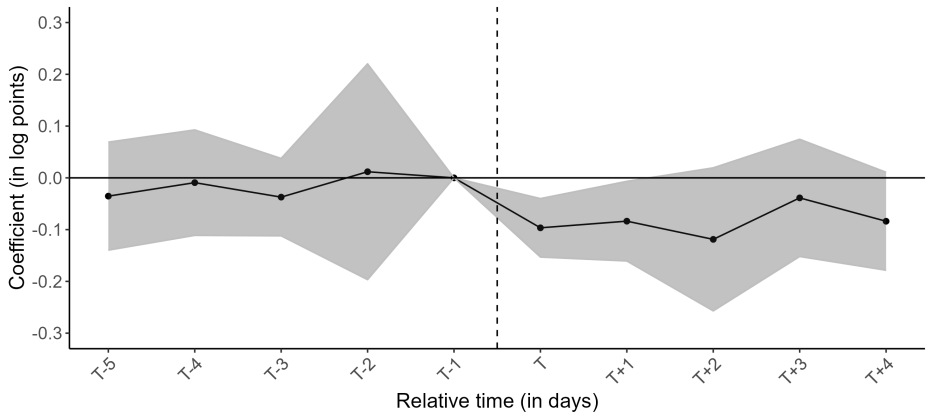


Figure 2.12 – Main results from event-study design using inverse page number weights.

Notes: This figure shows the results of the event-study regression of negative articles on the number of Universal Credit applications, weighted by the inverse page number.

the results from the main regression weighting by the inverse of the page number. The inverse is chosen because we want to give more weight to earlier articles, but beyond that, it is an arbitrary choice in the space of decreasing functions. Figure 2.12 shows the results from this exercise. Extracting precise quantitative meaning from the figure is complicated because of the weighting scheme, but it does qualitatively show that the results continue hold and that they seem to be larger (as compared to the main results in figure 2.11a). This therefore suggests that articles that appear earlier in the newspaper are driving our results, consistent with the hypothesis these articles are more likely to be read and/or more likely to be negative towards benefit seekers.

Positive and neutral stories We also consider the impact of positive and neutral articles on the number of applications for Universal Credit.[22] Figure 2.18 in the appendix shows the results for these events, which appear to have little effect on the number of applications.

4.3 Placebo test and Fischer randomisation inference

Given the low number of observations we are able to exploit, it is interesting to take another approach to inference by implementing a Fisher Randomization Test. This can play both the role of a placebo test and a way of generating an approximately exact test of the (sharp) null hypothesis.

[22]The methodology is identical to the one used above, replacing negative stories with positive or neutral stories. In particular, this implies that negative articles are not considered in the selection of events without overlap.

We implement this by randomly drawing the same number of events as we have originally and then applying the full procedure as described above (including restrictions on dates due to window overlap and bank holidays). We repeat this 10,000 times and plot histograms of the coefficients from the corresponding event-study regression. The results are shown in figure 2.13.

Table 2.4 – P-values and quantiles from randomisation inference

	T-4	T-3	T-2	T-1	T+1	T+2	T+3	T+4	T+5
p-value	0.960	0.633	0.464	0.970	0.104	0.039	0.145	0.719	0.283
original estimate	-0.001	0.013	0.019	-0.001	-0.039	-0.048	-0.037	-0.009	-0.030
2.5% quantile	-0.044	-0.046	-0.044	-0.045	-0.047	-0.048	-0.056	-0.057	-0.063
97.5% quantile	0.060	0.059	0.054	0.052	0.046	0.042	0.043	0.040	0.038

Notes: This table shows the result from 10,000 draws of randomly re-drawing events and computing the main event-study approach. This allows to compute p-values as $1 - \mathrm{E}[|\beta_o| > |\beta_{\mathrm{ri}}|]$ for each coefficient. It also shows the original estimate and the relevant 2.5% and 97.5% quantiles for direct comparison.

The figure shows a very clear pattern which confirms the results in the previous section. The levels of significance are slightly lower with $T + 1$ at just around 10% significance and $T + 2$ just around 5%. The p-values and corresponding quantile values are shown in table 2.4.

4.4 Robustness

We perform a battery of tests. Full results are given in appendix A.3.

Leave one out The results are robust to leaving out any one event. Figure 2.19 shows the coefficients for the two periods preceding and following the publication date. The results remain remarkably unchanged from the main specification. The number of applications two and three days before the publication is indistinguishable from the day before publication (the reference day) in all cases. The day of and the day after publication is lower than the day before publication by 4-5%.

Transformation of dependent variable Figure 2.20 shows the result for the main regression in levels rather than logs. The strong time trend means that earlier events get drowned out,

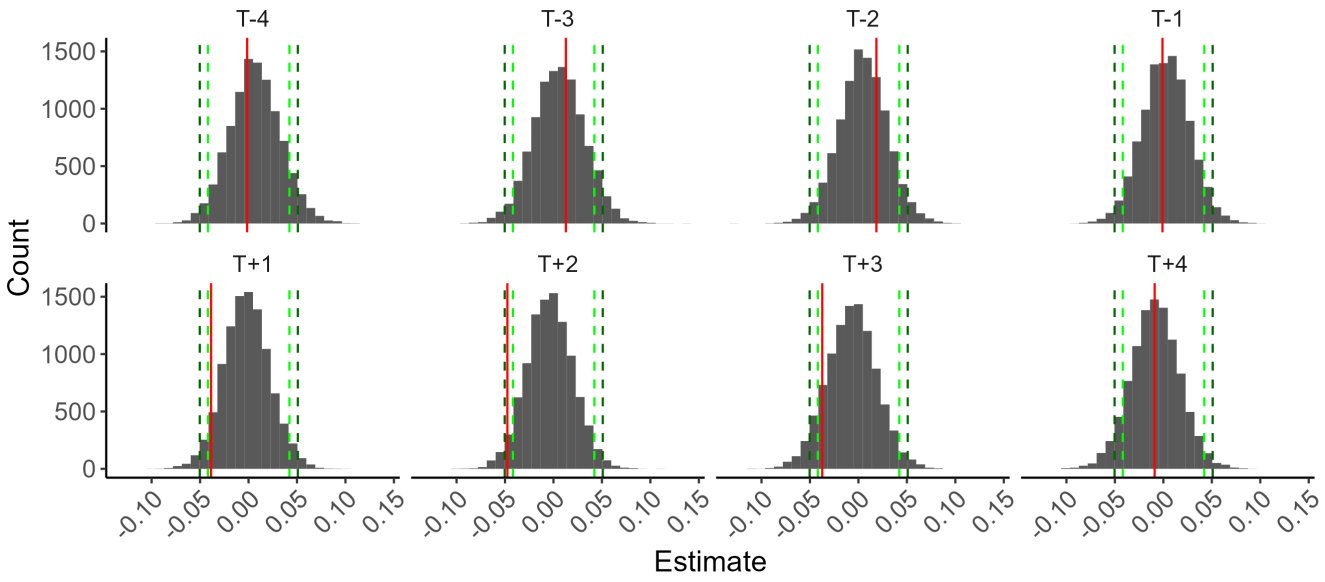


Figure 2.13 – Randomization inference for the main regression.

Notes: The event days are randomly re-drawn 10,000 times. The standard event-study approach discussed above consisting of residualising, restricting to non-overlapping windows and estimating an event-study specification is applied to every draw. The estimates for each draw are then plotted in the histograms for each relative time. The red line represents the original estimates. The light and dark green lines represent the corresponding 2-sided 10% and 5% quantiles.

meaning that this approach is not dissimilar to looking at the events in 2019 only. This increases the variance, to the point where none of the estimates are significant, but the figure exhibits a clear pattern of a drop in applications following a story.

Window length We explore the sensitivity of our results to the choice of window length. Figure 2.21 shows that results are nearly identical when taking a 12-day window, if slightly less significant, which is to be expected given fewer events (35 instead of 46).

Figure 2.22 shows the main results with a 6-day window (instead of 10 days). As expected, this makes the pre-trends more noisy and dampens the significance of the results, as there are now some treated days in the pre-period. Nonetheless, the patterns of the results are similar, particularly for the first income quartile and the fourth applications quartile.

Window type As discussed in section 4.1, we used a 'one-sided' removal of overlapping events: we removed an event if the event window of another event overlapped with its pre-period, but not its post-period. Here we consider 'two-sided' removal, where we remove all events with any

overlap. This reduces the number of events to 23 (compared to 46 in the main regression), and selects for the least important articles - those that received no follow-up coverage. Figure 2.23 shows the results of this specification. The magnitude of the effect is much smaller and lasts for a single day, but is nonetheless marginally significant at the 5% level.

Residualisation Two other options for the residualisation specification are a non-interacted specification:

$$\ln\left(\text{Applications}_t\right) = \text{dow}_t + \text{month}_t + \text{year}_t + \varepsilon_t \tag{2.2}$$

and a version which uses polynomial detrending:

$$\ln\left(\text{Applications}_t\right) = \text{dow}_t \times \left(\sum_{i=1}^{5} \gamma_i t^i\right) + \varepsilon_t \tag{2.3}$$

Appendix figures 2.15 and 2.16 show the equivalent of Figure 2.10 for these alternative specifications. Comparing these graphs shows why we choose the fully interacted specification for the main body of the text as it performs significantly better at residualising the irregular time trend in the data.[23] Nonetheless, as a robustness check, we show that the results for the different residualisations are nearly identical (see section 4.4).

In figures 2.24 and 2.25, we show the results when choosing a non-interacted residualisation specification as described in equation (2.2) or a polynomial residualisation specification as described in equation (2.3) respectively.

The conclusions are identical to the main specification. The treatment effects are more significant but the pre-trends are more noisy, which is consistent with a less good residualisation.

Dropping Saturdays Another concern mentioned above is that in the case of imperfect residualisation, events on Saturdays could be driving the results because of the drop in applications that occurs on Saturdays on average, even without a treatment event. In figure 2.26, we therefore show the results from the main specification when dropping events that occur on Saturdays.

As can be seen from figure 2.9, this robustness test is very stringent as it drops a significant fraction of events (we are left with 33 events in this case), we therefore expect the results to be

[23] One concern with the fully interacted specification using fixed-effects is that these might induce discontinuities. This concern is already addressed through the polynomial specification. We also check that dropping events that are at the very beginning or end of the month does not affect the results (results not shown).

quite noisy. The figures are consistent with a very similar pattern as for the main result, but significantly more noisy.

5 Discussion

So far we have established a causal relationship between the publication of stories containing negative content about benefit recipients and a decline in the number of applications for Universal Credit. In this section we present a conceptual framework which models stigma as a combination of social-image and self-image concerns. We explore our results in light of this conceptual framework, discussing stigma and non-stigma mechanisms that might lead to a change in the number of applications following a negative *Sun* story. Other mechanisms cannot account for this effect. We particularly focus on the short-term nature of our observed effect. We then use our conceptual framework to give a back-of-the envelope cost of stigma in this context.

5.1 Conceptual Framework

How does negative media content affect benefit take-up? How are stigma, social-image and self-image concerns relevant in this context? A large body of literature in economics, sociology and psychology has considered how self-image, social image and stigma affect behaviour. Here we draw on this to conceptualise benefit stigma. This is not intended to be a model, merely an exposition.

Benefits-eligible individual i makes a take-up decision $d_i \in \{0, 1\}$. This decision may be observed by some reference group j. We assume there is a single reference group, although the framework can be extended to multiple groups.[24] If the individual takes up the benefit, they are subject to some social-image cost and some self-image cost. We conceptualise the stigma cost as the sum of the social-image cost and the self-image cost. Both of these costs are explored further below.[25] Take-up also confers other benefits, B_i, (eg. income) as well as costs C_i (eg. time taken to apply). They make their take-up decision based on whether the benefit outweighs these costs.

[24]It is common in the literature on non-take-up of social benefits to distinguish between social stigma and institutional stigma. Here these types of stigma can be modelled with different reference groups.

[25]This is equivalent to assuming that social and self-image perfect substitutes, which we use for the sake of simplicity in this basic framework, and does not affect our analysis. However, they may be substitutes or complements; higher social image may compensate for lower self-image, or it may be necessary to have a certain level of social image before a high self-image can be attained (Bursztyn and Jensen, 2017).

$$d_i = \begin{cases} 1, & \text{if } B_i - C_i - \text{Social}_{ij} - \text{Self}_{ij} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

5.1.1 Social image

Here we rely heavily on Bursztyn and Jensen (2017) (who in turn draw on Bénabou and Tirole, 2006). Making decision $d_i \in \{0, 1\}$ might reveal information about i's type, t_i. There are many different 'bad' types that benefit recipients have been associated with, including low ability, free rider (Friedrichsen et al., 2018), poor (Holford, 2015), and fraudulent (Gavin, 2021). Here we assume a singular 'bad' type, $t_i \in \{t_B, t_G\}$, as representative of these, but the framework could be extended to model different types.[26]

We define the social image term in i's take up decision as:

$$\text{Social}_{ij} = \lambda_i \cdot e_i\left(\omega_{j,t_B}\right) \cdot \left(e_j[t_i = t_B \mid d_i = 1] - e_j[t_i = t_B \mid d_i = 0]\right)$$

ω_{j,t_B} represents the social penalty of being seen by group j as the 'bad' type, t_B. However what matters for i is their perception of this penalty rather than the penalty itself. Thus $e_i\left(\omega_{j,t_B}\right)$ represents i's belief of the social penalty. λ_i represents how much i cares about receiving this social penalty. The final term is the probability that i will receive the social penalty. More specifically, i's take-up decision might be imperfectly observed, so the reference group forms a belief about i's decision, $e_j\left(d_i \mid d_i = 1\right)$, based on some potential available evidence. Conditional on this, they form an expectation of i's type. Social image depends on the difference between this expectation, and the expectation in absence of the take-up signal. We do not model second order concerns over this probability, such as i's beliefs about the likelihood of their take-up decision being observed and their beliefs about how this affects j's perception of their type, but these could also play a role.

If i's take-up decision is completely unobserved or it gives no information about i's type $(e_j(d_i \mid d_i = 1) = e_j(d_i \mid d_i = 0))$, then this term equals 0 and so the social image term drops out. In our context, take-up decisions are not perfectly observed. Applications can only be made

[26] In this set up, being a benefit recipient itself is not the reason for social or self judgement, but instead the signal that this gives about the likelihood of i being a different bad type. This seems a reasonable framework in this context, as survey evidence shows a large majority of people find at least some subgroup of benefit recipients 'deserving' (Geiger, 2021). This may contrast with other types of social judgements, such as homophobia.

online and payments are usually by direct debit. However, anecdotal evidence suggests that people tend to know when others around them are claiming benefits, meaning individuals have a reasonable expectation that their take-up decision would be suspected if not directly observed.

5.1.2 Self-image

Following Bénabou et al. (2018), we model self-image analogously to social image, with the reference group being i themself.

In the case of self-image, it might be thought that an individual already knows their type. However, if we consider types such as "lazy" or "undeserving", it is easy for us to see that individuals may not have true knowledge of their own type. The literature on self-signalling (Bodner and Prelec, 2002; Mijovic-Prelec and Prelec, 2010), divides the self into an "actor self" who makes the decision and a "judge self", who derives utility from their interpretation of the action. Then self-image can be modelled as:

$$\text{Self}_i = \mu_i \cdot (e_i[t_i = t_B \mid d_i = 1] - e_i[t_i = t_B \mid d_i = 0])$$

The final term represents the probability that i themself (the 'judge') see i as the 'bad' type. In other words, it represents how strongly individuals link the action with being 'bad' type. In this context, the take-up decision is perfectly observed. Further the penalty for perceiving themself as the 'bad' type and the amount i cares about receiving this penalty are rolled into a single term, μ_i (as both are determined by the 'judge' self). μ_i may be a function of i's belief of the social penalty, $e_i\left(\omega_{j,t_B}\right)$ (Ghosal et al., 2022; Goffman, 1986). This is different to social image concerns. In a case where i knows for certain that their take-up decision will be unobserved, but also knows for certain that being the 'bad' type is associated with a large social penalty, this may increase their self-image cost, despite the social-image cost being 0. This is commonly known as self stigma, or internalised stigma.

5.2 Mechanisms

5.2.1 What happens when a *Sun* story is published?

Negative stories may affect both social and self-image. Specifically, a negative story can potentially impact $e_i(\omega_{j,t_B})$, an individual's belief about the social penalty for being seen as a 'bad' type, by increasing the saliency of social judgements of this type. This would increase the social-image cost and therefore disincentivise application. We also earlier argued that the individual's weight on self-image, μ_i is increasing in $e_i(\omega_{j,t_B})$, so this would also increase the self-image cost. A negative article could additionally change ω_{j,t_B} itself, which would have a similar impact.

Such an article may also change $e_j[t_i = t_B \mid d_i = 1]$ (or the self-image analogue), by increasing the association between receipt of a social benefit and 'bad' types. In particular, many of the articles in our sample associate benefit receipt with fraud, extremely large families, obesity and other types that may be seen as 'bad' by society at large. If an article increases the social (or self) association between receipt of benefits and one of these types, then this will increase the social- (or self-) image costs and disincentivise applications. There may also, or instead, be second order effects on this term, if an individual believes that the probability of being associated with a 'bad' type has increased, or they believe that their action is more likely to be observed.

In general we cannot distinguish between these sub-effects. It seems likely multiple effects are at play and that they interact with each other in increasing the stigma cost. Nonetheless we believe it is most likely that the effect is driven through the $e_i(\omega_{j,t_B})$ term. Our identification strategy means that we identify the effect of a story after a period of at least ten days without negative content about benefit seekers. The stories act as "reminders" of society's negative view of benefit claimants. Individual stories may change ω_{j,t_B} and $e_j[t_i = t_B \mid d_i = 1]$ over time, but in the short-run, where we are able to identify the effect, it seems more likely that stories affect the saliency of underlying social parameters, rather than changing the social parameters themselves.

Institutional stigma We can, however, rule out institutional stigma. This can be seen as social stigma where j is the institution, or more generally as the "feeling of disrespect from the process of claiming benefits" (Geiger et al., 2012). In our set-up, the effect that we observe is unlikely to be primarily driven by institutional stigma. Individuals apply online and have no further contact with the institution on the day of application. They have no direct contact with any individual

working for the institution. Therefore whether an article is published or not, $e_i\left(\omega_{institution,t_B}\right) = 0$ and so the institutional social image cost drops out.

Benefit fraud Many of the stories in our sample contain information about benefit fraud (see figure 2.14). Given this, a sensible question to ask is why these would be stigmatising of (non-fraudulently) taking up benefits. Our conceptual framework helps us to shed light on this issue. These types of articles can increase both the social and self-image cost of claiming benefits, if we view "fraudulent" as a 'bad' type. First, articles about people who are claiming benefits fraudulently increase $e_j[t_i = t_B \mid d_i = 1]$, the probability that the reference group think that an individual is 'bad' type (or the individual's perception of this probability), increasing the social-image cost of applying. The perceived rate of fraudulent claims is far higher than the actual level.[27]

Second, these stories may increase the social-image cost, by reinforcing the idea that the type of person who claims benefits is also the type of person who is capable of criminal behaviour, thus increasing μ_i, the individual's concern about seeing themself as 'bad' type.

5.2.2 Confounders

Our conceptual framework also helps us to see what kind of mechanisms would stop us from being able to interpret our results as an effect of stigma. In particular, anything that increases other benefits, B_i, or decreases others costs, C_i of applying will confound our interpretation. Here we consider various mechanisms by which newspaper stories might affect B_i or C_i, but find that none of them can account for the effect that we observe.

Information effect Negative articles about benefit seekers also contain information about benefits themselves. Therefore they may make people aware of benefits that they did not know about, or aware of their eligibility for benefits. Lack of information about benefits has been found to be a driver of non-take-up in previous work (Bhargava and Manoli, 2015). However, if an information effect is present, we would expect this to increase B_i, and therefore increase the

[27] In 2016, 22% of respondents to the British Social Attitudes Survey thought that the majority (ie. greater than 50%) of claimants of Unemployment Benefits were doing so fraudulently, which is more than ten times the actual level of fraud. See https://www.bsa.natcen.ac.uk/media/39144/bsa34_benefit_tax_final.pdf, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664827/fraud-and-error-stats-release-2016-17-final-estimates.pdf

number of applications after the publication of one of these stories, rather than decrease it. If anything, this means that our estimated effect is a lower bound on the stigmatising effect of the stories.

Complexity effect Newspaper stories may also include information about the complexity of applying for benefits. This could include information on the process itself, or information on issues with the rollout of Universal Credit. This could potentially increase individuals' perceptions of C_i, and therefore decrease their likelihood of applying for Universal Credit. Finkelstein and Notowidigdo (2019) found that complexity costs can reduce benefit take-up. To rule out this mechanism, we looked for content which discussed the processes of application, or problems with the Universal Credit rollout among our sample of negative articles. We found that none of the articles touched on these topics at all. Therefore it seems unlikely that our effect is driven by an increase is perceived complexity.

Benefit fraud As mentioned earlier, many of the stories in our sample focus on cases of benefit fraud. Therefore another possible alternative mechanism for the effect of stories on application behaviour is that these stories deter those people from applying who were planning on committing benefit fraud, arguably increasing C_i. However, only 3.1% of Universal Credit expenditure was due to fraud in 2016, with the vast majority of this being due to a failure to notify a change in circumstances, rather than due to making an initially fraudulent application. So even if 100% of people who were planning on applying fraudulently were exposed to the *Sun* story, and every single one of them was deterred from applying due to the story, this mechanism could only account for a fraction of our observed effect.

5.3 Delaying or detering

An important question about our results is whether they are indicative of a short-term substitution effect where individuals delay their benefit claim by a couple of days (or weeks), or of a definite non-take-up where the affected individuals never end up claiming. The short-term identification strategy we employ in this paper makes answering this question complicated, as we are only able to see what happens to the number of applications in the short time period around the publication of a negative newspaper article. What we do observe, however, is that

treated individuals do not appear to apply in the first five days after the publication of a negative article. Importantly, our results should not be interpreted as saying that individuals delay their application for social benefits by four days. Application numbers return to baseline levels after five days, and we do not see an increase in applications, which we would observe if individuals were simply shifting their application date by a few days. A better interpretation in that the treatment period lasts four days. Individuals who would have applied in the four days following the publication of a story appear to adjust their behaviour in response, whereas individuals who would have applied on the fifth day do not appear to adjust their behaviour. Therefore it seems that a story can be associated with four days of treatment, on average, with individuals on the fifth day not being treated. This interpretation further supports the saliency mechanism described above. Plausibly, after a four-day period, the increased saliency of the social penalty, $e_i\left(\omega_{j,t_B}\right)$, dissipates, as other news stories or topics become more salient.

This does not preclude a delay in take-up in the more medium-run of a couple of weeks. However, this is not unusual with any determinant of non-take-up. Much of non-take-up comprises delaying rather than never claiming, and procrastination has been shown to be a significant determinant of non-take-up (Dynarski, 2007; Sunstein, 2013; Narayan, 2020). Further, as demonstrated below, even individuals do later take up social benefits, they are still forgoing a substantial sum of money.

5.4 Climate of stigma

As well as increasing saliency, over time newspaper articles with negative coverage of benefit recipients may also create or reinforce a climate of benefit stigma. This could be seen as a larger, long-run shift in the social penalty, ω_{j,t_B}, beyond the short run fluctuations that could be one of the mechanisms for the effect observed in this paper.

Our approach cannot speak to these longer-run effects. In this respect, the effect we document is only a lower bound of the effect of stigma on benefit take-up. Work identifying the climate of stigma around welfare would be an interesting avenue for future research.

5.5 Magnitude of effect

We document a 4-5% reduction in the number of applications to Universal Credit over at least the three days after the publication of a negative story about benefit recipients. Figures on non-take-up of Universal Credit are not published, but it may be somewhere between 15% and 40%, based on rates for the legacy benefits that it replaced (see figure 2.4). If we take the midpoint of this, then around 72.5% of untreated eligible people apply. If 72.5% of people apply after a 5% reduction in take-up, then 76.2% ($= 72.5/0.95$) would have applied beforehand, meaning a 3.7 percentage point drop in take-up. This means we can explain somewhere around 13.5% of the baseline non-take-up (27.5%) on treated days. Over the period, we have 518 'treated' days (days where a negative article is published or the subsequent two days), which is 23.7% of total days, meaning overall we explain around 3.2% of non-take-up.

The size of this effect is comparable, if a little smaller, to that which has been found in work looking at other drivers of non-take-up. Finkelstein and Notowidigdo (2019) find they are able to reduce non-take-up of SNAP by 5% with an information intervention and by an additional 7% with an intervention to reduce complexity. Bhargava and Manoli (2015) find a reduction in take-up of 22% from an information intervention. However, the effect that we document is an extremely conservative estimate of the overall effect of stigma on non-take-up, as we are only looking at one vector of stigmatisation, a single newspaper, and we are only able to identify the short-run effect, rather than the climate of stigma. Friedrichsen et al. (2018) look at the effects of stigma on take-up in the lab and find a far larger effect than documented here; they find that it explains around 71% of non-take-up in their sample.

As the receipt of social benefits has a monetary value attached, we can approximate a back-of-the-envelope computation of how much individuals on average are willing to pay to avoid stigma, without having to impute a value.

The minimal amount of monthly Universal Credit payment in 2021 was £368 (€429, $462) for a single individual, yielding a daily allowance of around £12 €14, $15).[28] If we conservatively assume that a story only delays take-up by three days, the cost of a single stigmatising story is £36 €42, $45). This is a lower bound, as mentioned above, we see no evidence that individuals do in fact apply in the medium run. As an upper bound, if a story fully deters the affected individuals from taking-up, then the cost would go up to £810 €944, $1017), which is the average

[28] see for instance `https://www.gov.uk/universal-credit/what-youll-get`

Universal Credit paid out in 2021[29].

These are rough estimates, but even the lower bound shows that the form of stigmatisation we document in this paper has a high cost to individuals. Indeed, our estimate is significantly bigger than for instance Dellavigna et al. (2017) who find that individuals have "a value of voting "to tell others" of about $15".

5.6 Stigma as a targeting mechanism

It has been suggested that stigma can improve the targeting of benefits to those who most need them most (Besley and Coate, 1992). Our results indirectly provide evidence against this hypothesis. In particular, the fact that the decrease in applications is strongest in the lowest income regions suggests that stigma deters the poorest from taking-up benefits. This conclusion is in line with other recent results (Deshpande and Li, 2019; Homonoff and Somerville, 2021).

6 Conclusion

This paper documents the effect of individual news stories, containing negative information about benefit recipients, on the take-up of social benefits.

We use high frequency data on the number of applications to a "catch-all" working-age means-tested benefit in England, Universal Credit, and conduct an event study around the period of publication of a negative news story in *The Sun*, a highly-read tabloid newspaper.

We find a 4-5% drop in the number of applications in the three days following the publication of such stories. We find no evidence of a compensating increase in applications after these three days, in the short run. This effect is more pronounced in areas with low average income and high base levels of benefit applications.

This effect cannot be explained by information effects, deterring benefit fraud, or increasing institutional stigma. Instead, it seems to be driven by increasing social and personal stigma.

This result provides evidence that stigma has a role in shaping take-up behaviour, as well as opening the 'black box' of media persuasion, by showing the effect of individual stories on economic behaviour.

[29] See `https://www.gov.uk/government/statistics/universal-credit-statistics-29-april-2013-to-1` `universal-credit-statistics-29-april-2013-to-14-july-2022`

A Appendix

A.1 Examples of negative coverage of benefit recipients

Disability swindler is busted at fit club (2014-01-11)

A BENEFIT cheat who swindled £21,000 claiming she could barely walk was caught taking part in FITNESS classes. Anne Bird, who has osteoarthritis, was filmed by undercover investigators after a tip-off. The 60-year-old, who said she needed help dressing, did pilates, aerobics and spinning bike classes, a court heard. She was also filmed doing sit-ups at a gym. She told the Department for Work and Pensions she needed the higher rate of disability allowance because she also had problems with her knees and hips. Despite Bird, from Pleck, West Mids, being entitled to some benefits, between October 2008 and August 2012 she was overpaid £21,109. She admitted two counts of making a false declaration on forms at Walsall magistrates. Bird got 12 weeks jail suspended for 12 months.

THE Sun SAYS Addict clamp (2015-02-14)

FEW would argue against sickness benefit claimants losing their cash if they refuse help for treatable health problems. But life's not always that cut and dried. It seems logical to strip welfare from jobless alcoholics, junkies and the obese if they could sort out their lives but can't be bothered. In some cases that will be justified. But many drink and drug addicts aren't thinking straight or capable of seeking treatment. Making them penniless won't help. We all want them back on the straight and narrow and in work. The Government needs to be careful how much brute force they use.

Benefits cap calls (2016-07-09)

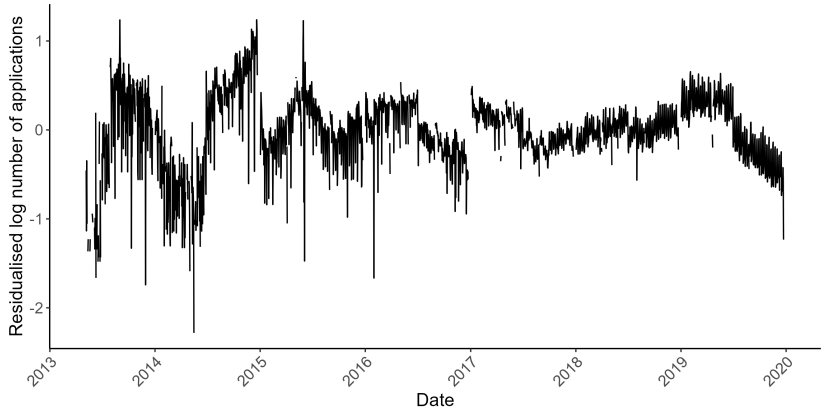
A PETITION demanding a benefits cut for "Britain's most shameless mum" has topped 15,000 signatures. Mum-of-12 Cheryl Prudham, 33, reportedly rakes in £40,800 in handouts. But the part-time cleaner, of Lancashire, said on telly she is saving for a boob job and wants baby 13 with a sperm donor after dumping her hubby. It prompted critics to set up a change.org petition to get her cash cut.

GUT TO GET A JOB, Help dole fatties back to work, says report (2016-12-06)

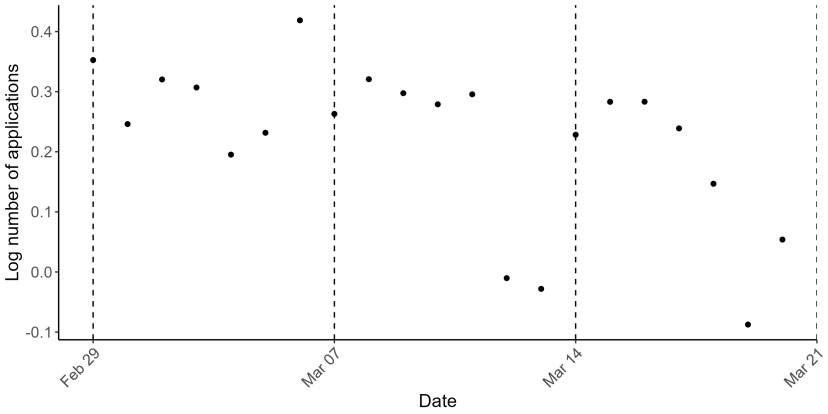
OBESE people on the dole should be told to see their GP to help get them back to work. A report also suggests Jobcentre staff should direct them to slimming clubs. The proposals are part of an official review into the effects of obesity and drug and drink addiction. Author Dame Carol Black warned fat jobseekers were ignored by potential bosses. Her report said 1,600 people who claim Employment and Support Allowance (ESA) were recorded as disabled because of their severe obesity. But up to 800,000 ESA claimants had disabilities where their weight may have been a contributing factor. The 140-page Department for Work and Pensions report stated that because almost a quarter of working-age adults were obese, many were in work. But there were fewer heavily overweight people in employment than those of a “normal" weight, and that gap soared for the severely obese. Dame Carol urged ministers to investigate the effects of obesity on the working population. Last night Downing Street welcomed the report but distanced the Government from making any benefit sanctions against the disabled.

Figure 2.14 – Examples of articles that we classified as containing negative content about benefit recipients

A.2 Other figures



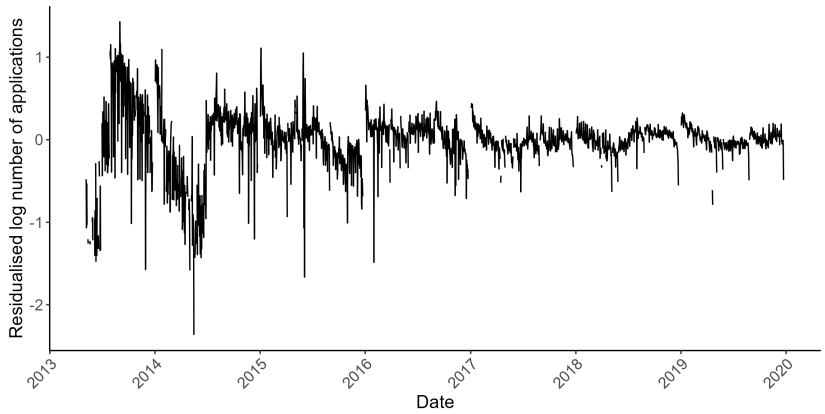
(a) Daily Universal Credit applications, 2013-2019, after the alternative residualisation without interactions.



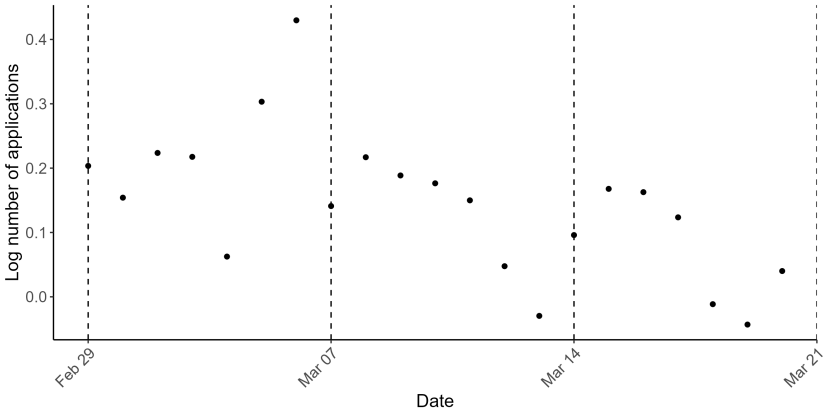
(b) Daily Universal Credit application over the 3 weeks between February 29th and March 20th 2016, after the alternative residualisation without interactions. The dashed lines are Mondays.

Figure 2.15 – Daily Universal Credit applications after the alternative residualisation without interactions.

Notes: Panel 2.15a shows the overall trend of the log of the number of applications after residualising using equation (2.2) in the main text. Panel 2.15b shows a randomly chosen three-week interval between February 29th and March 20th 2016, after the same residualisation. This can be compared with Panels 2.10a and 2.10b to compare this residualisation choice with the one in the main body of the text.



(a) Daily Universal Credit applications, 2013-2019, after the alternative residualisation using polynomial detrending.



(b) Daily Universal Credit application over the 3 weeks between February 29th and March 20th 2016, after the alternative residualisation using polynomial detrending. The dashed lines are Mondays.

Figure 2.16 – Daily Universal Credit applications after the alternative residualisation using polynomial detrending.

Notes: Panel 2.15a shows the overall trend of the log of the number of applications after residualising using the specification with polynomial detrending (equation (2.3)). Panel 2.15b shows a randomly chosen three-week interval between February 29th and March 20th 2016, after the same residualisation. This can be compared with Panels 2.10a and 2.10b to compare this residualisation choice with the one in the main body of the text.

A.3 Results of robustness tests

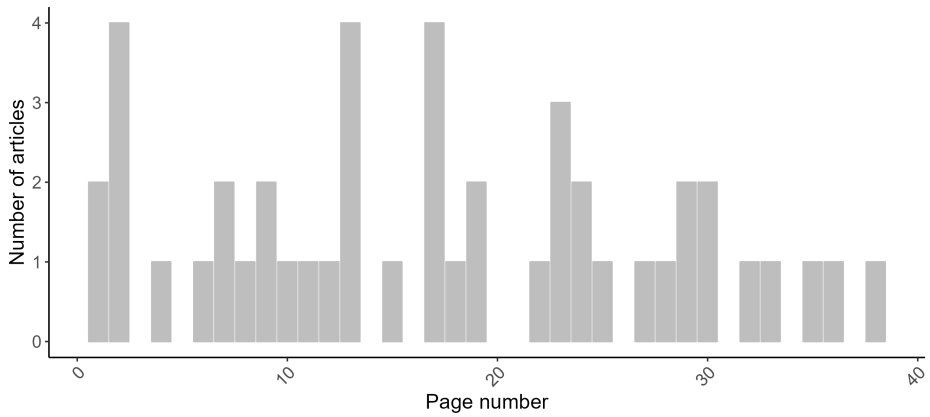
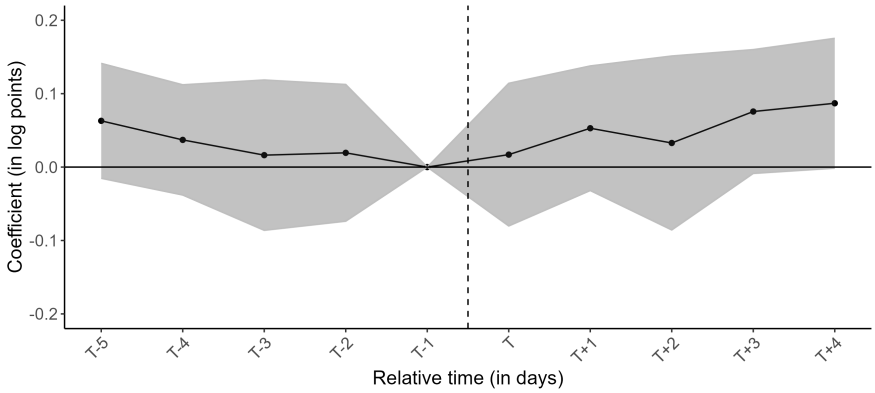
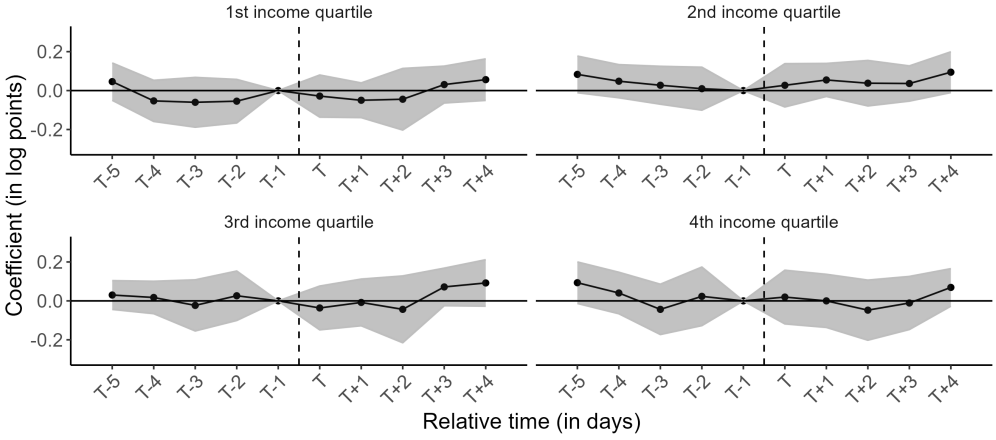


Figure 2.17 – Distribution of page numbers for the selected events for the main regression.

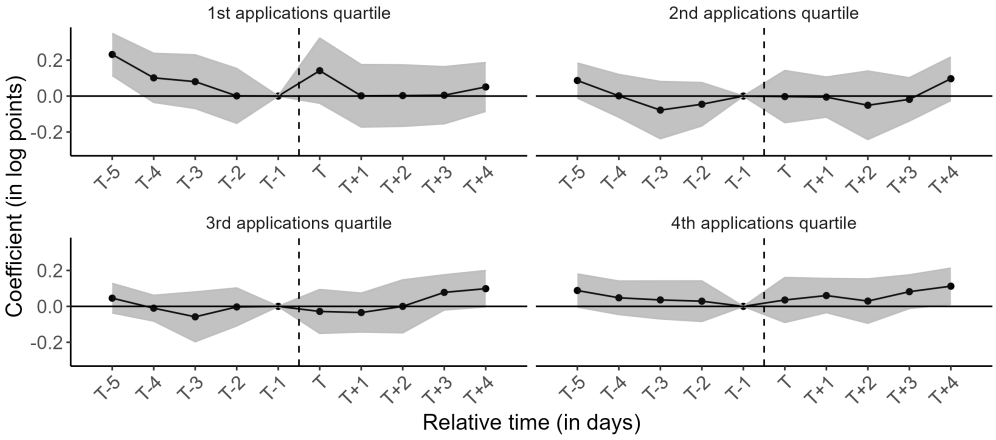
Notes: This figure shows the distribution of page numbers for the subset of events that are selected for the main regression.



(a) Main regression



(b) Heterogeneity by regional quartile of income



(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.18 – Results for positive or neutral stories from event-study design.

Notes: Panel 2.18a shows the results of the event-study regression of positive or neutral articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 10-day windows, with only one-sided overlap between events allowed. Panels 2.18b and 2.18c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.

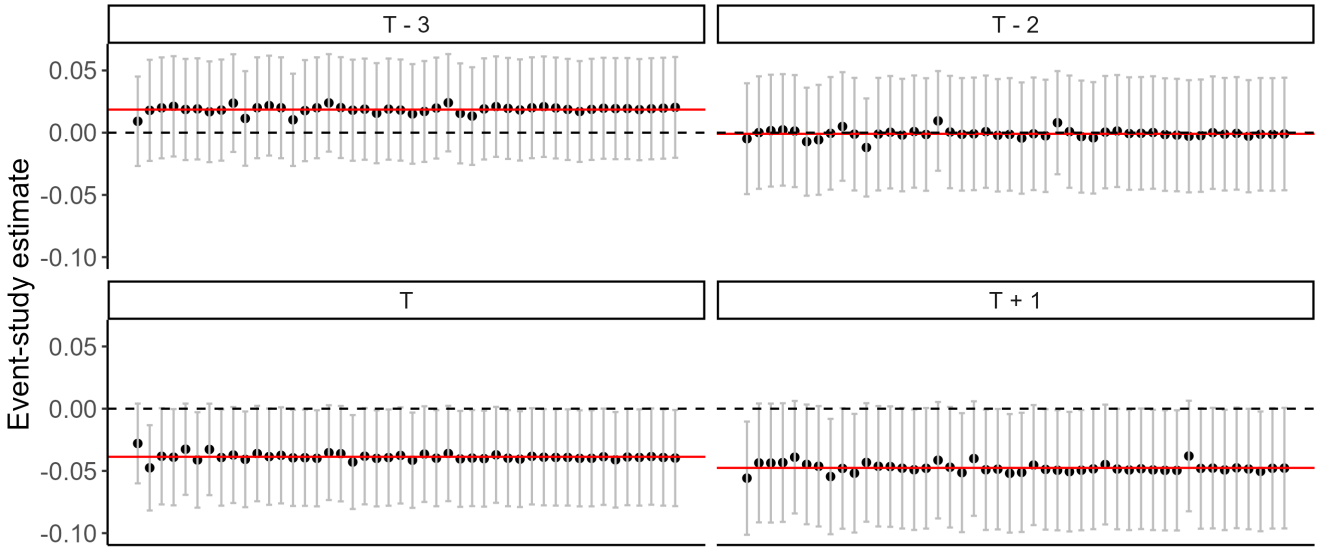


Figure 2.19 – Leave-one-out analysis

Notes: The graph shows the coefficients in the main regression, leaving each event out one at a time. The red line represents the coefficient in the main regression.

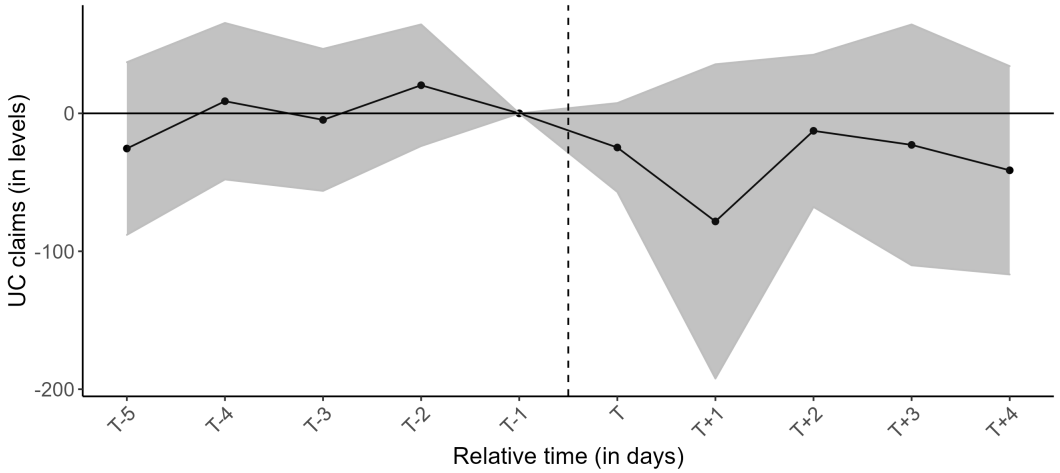
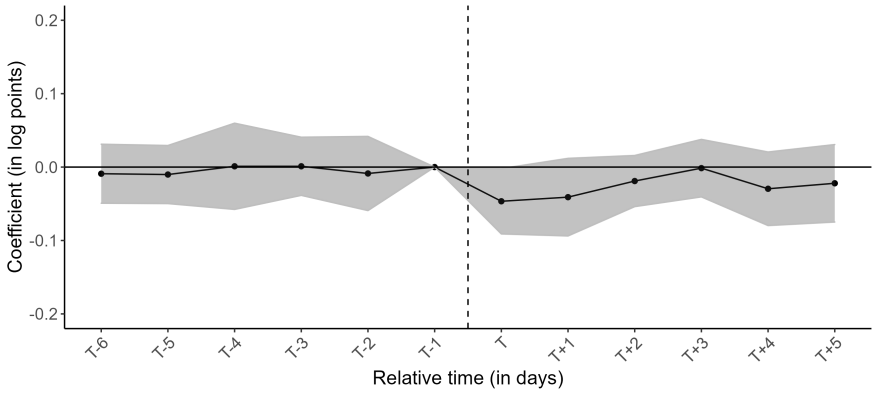
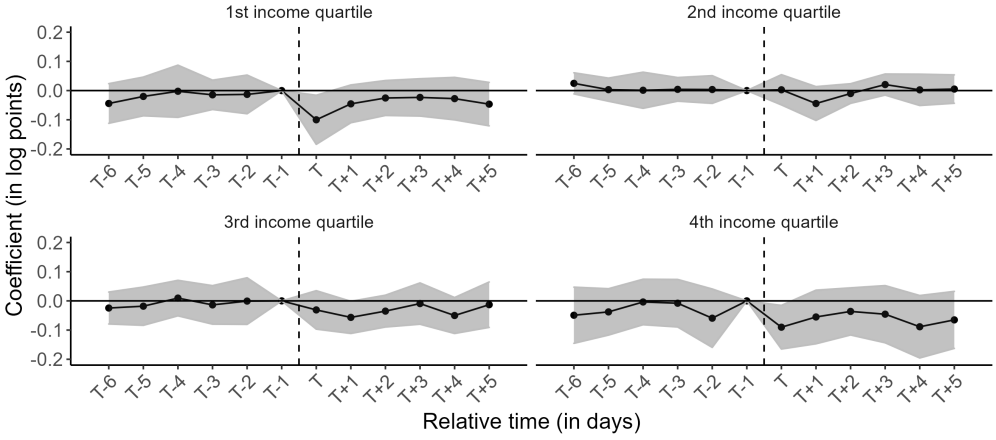


Figure 2.20 – Main event-study in levels

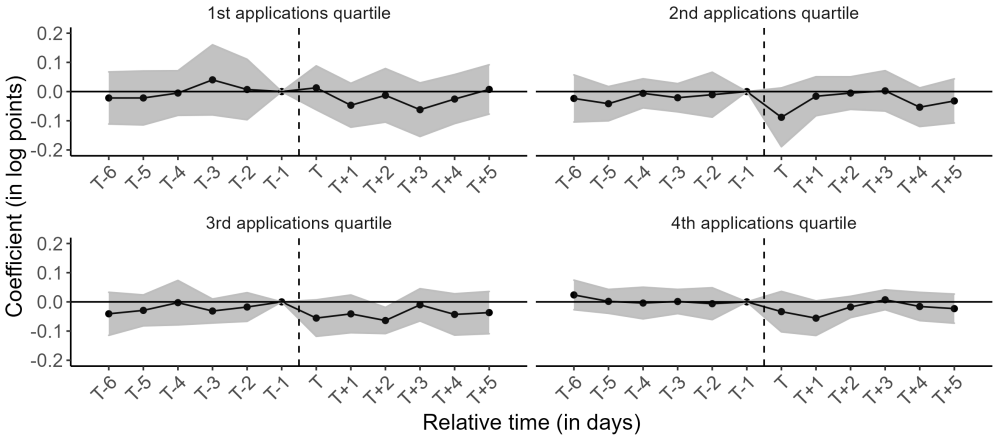
Notes: The graph shows shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 10-day windows, with only one-sided overlap between events allowed.



(a) Main regression



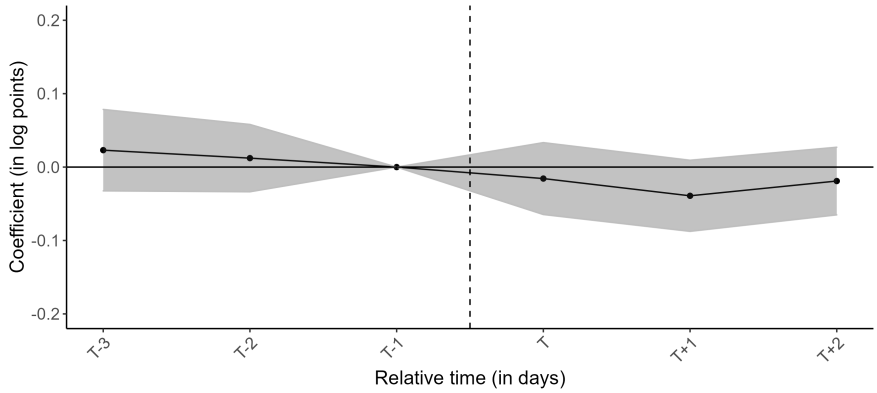
(b) Heterogeneity by regional quartile of income



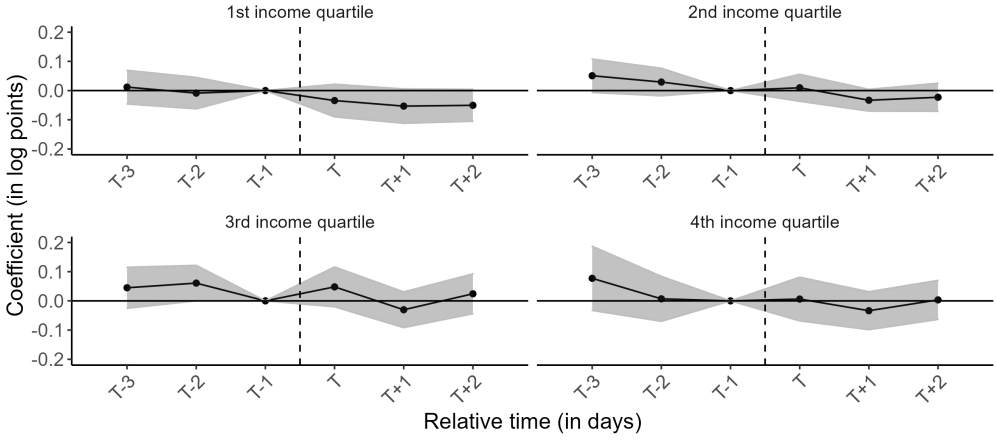
(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.21 – Main results from event-study design with a 12-day window

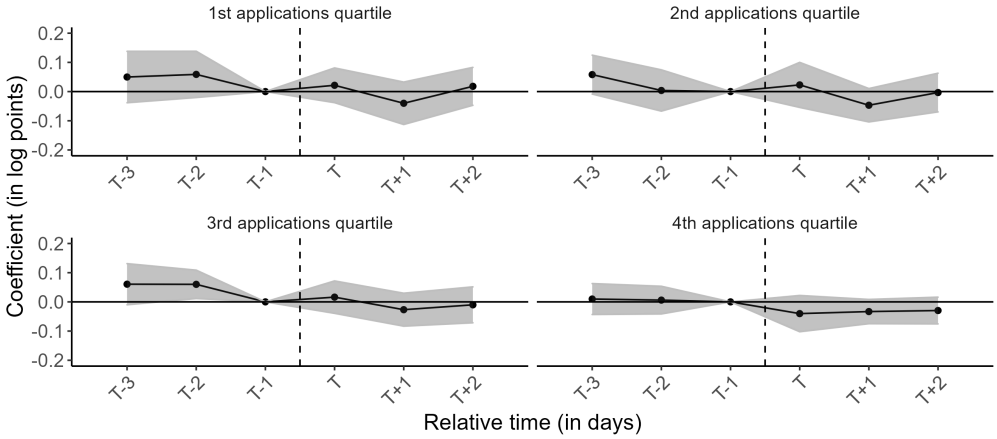
Notes: Panel 2.21a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 12-day windows, with only one-sided overlap between events allowed. Panels 2.21b and 2.21c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.



(a) Main regression



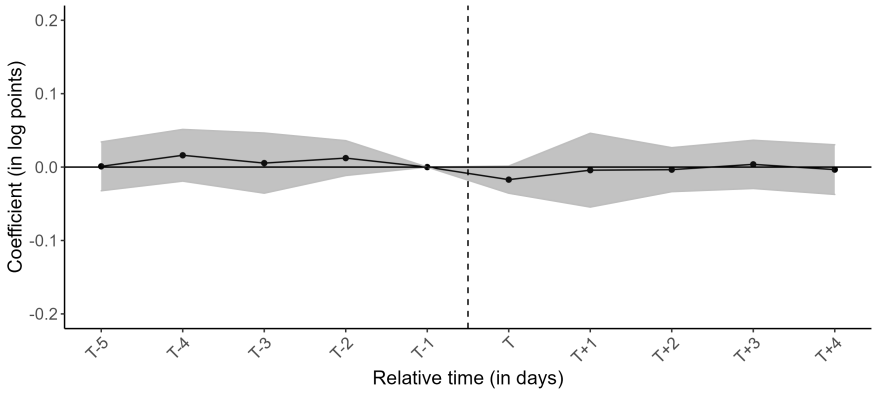
(b) Heterogeneity by regional quartile of income



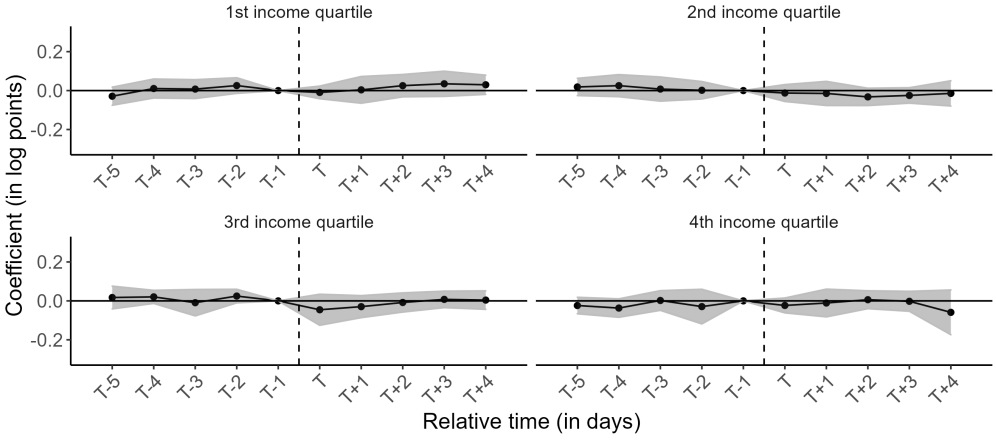
(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.22 – Main results from event-study design with a 6-day window

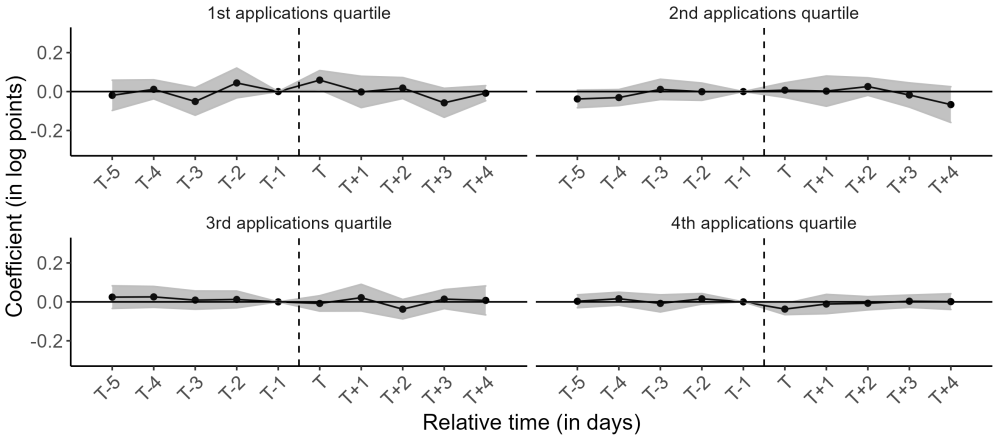
Notes: Panel 2.22a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 6-day windows, with only one-sided overlap between events allowed. Panels 2.22b and 2.22c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.



(a) Main regression



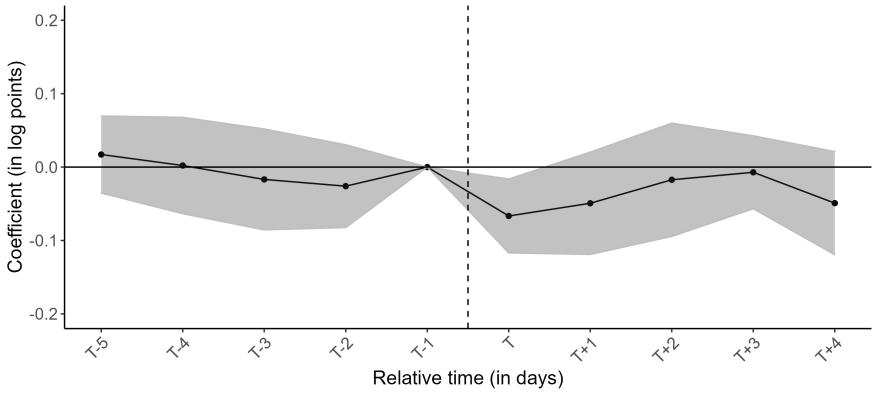
(b) Heterogeneity by regional quartile of income



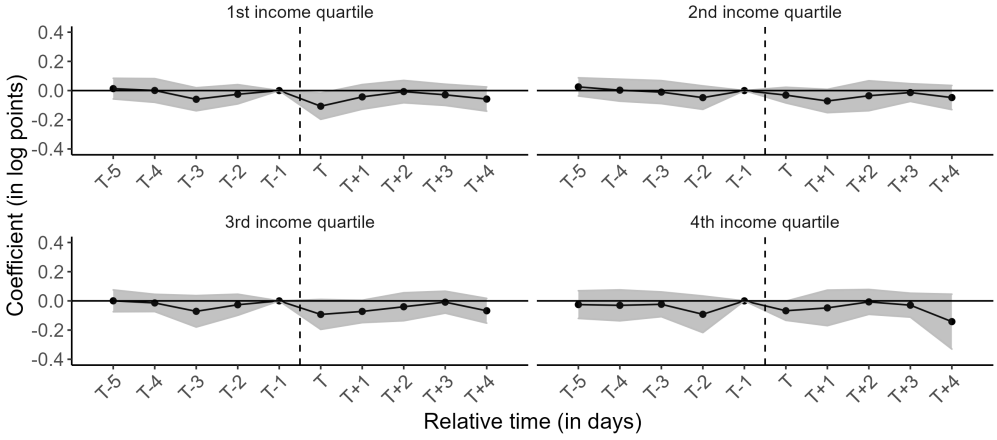
(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.23 – Main results from event-study design with two-sided 10-day window.

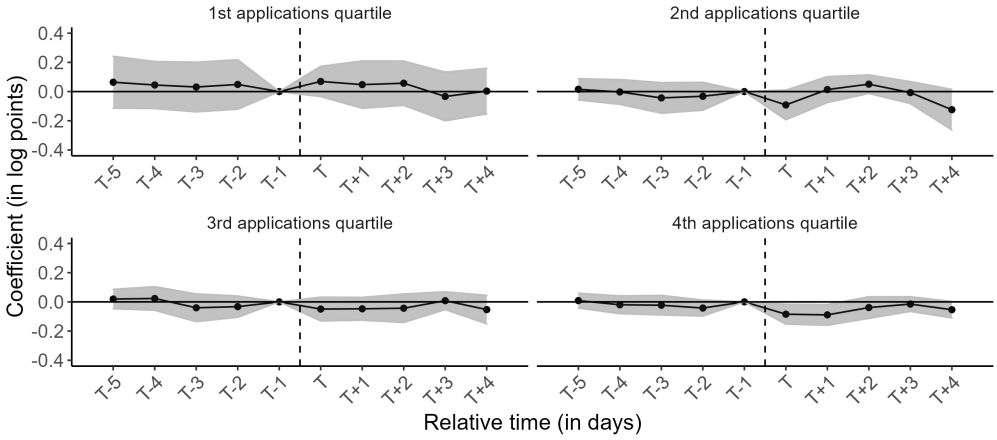
Notes: Panel 2.23a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 10-day windows, with only no overlap between events allowed. Panels 2.23b and 2.23c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.



(a) Main regression



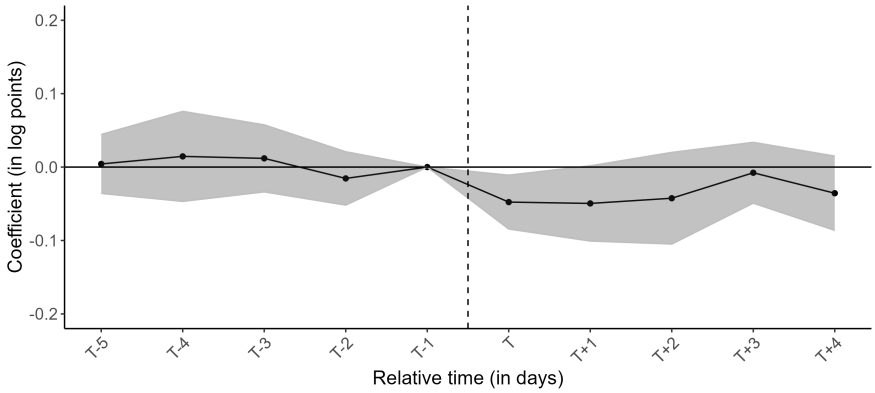
(b) Heterogeneity by regional quartile of income



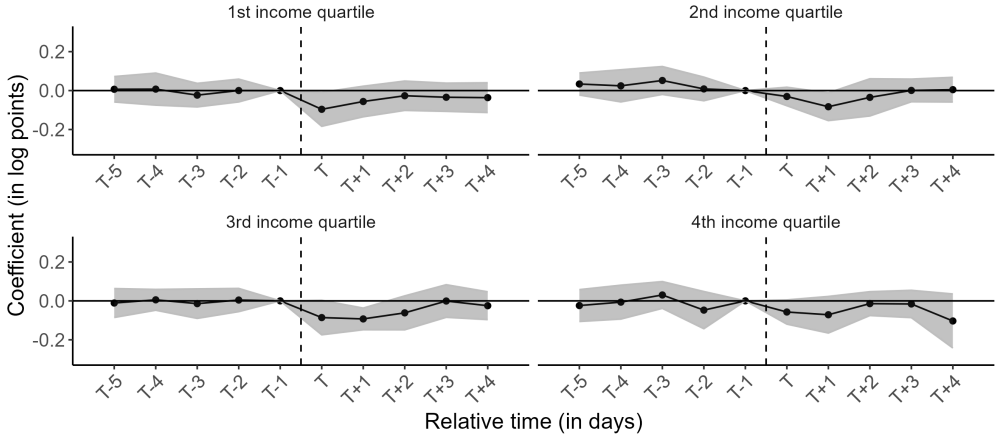
(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.24 – Main results from event-study design with residualisation specification without interaction

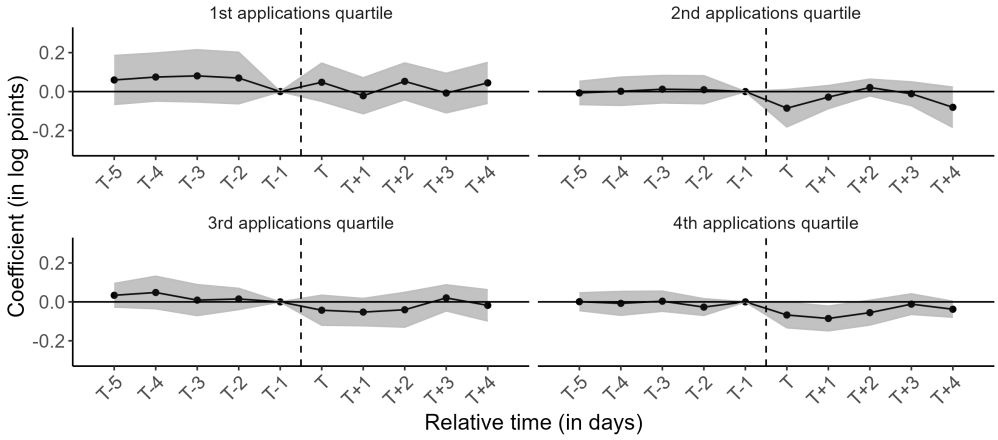
Notes: Panel 2.24a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week, month and year fixed-effects (without interaction) and then restricted to 10-day windows, with only one-sided overlap between events allowed. Panels 2.24b and 2.24c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.



(a) Main regression



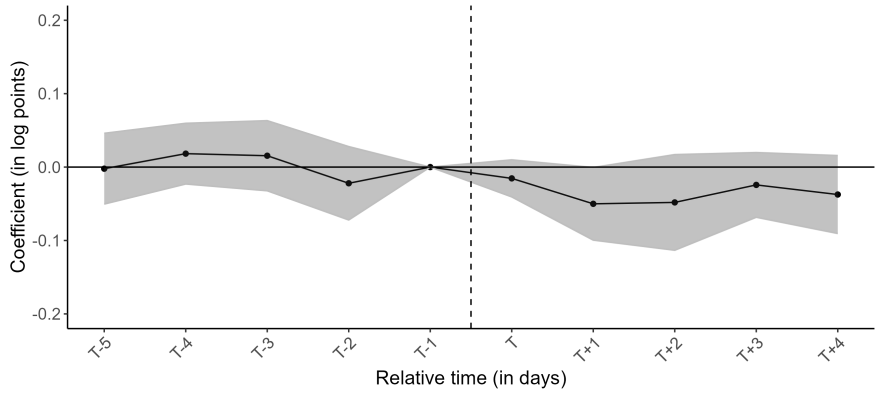
(b) Heterogeneity by regional quartile of income



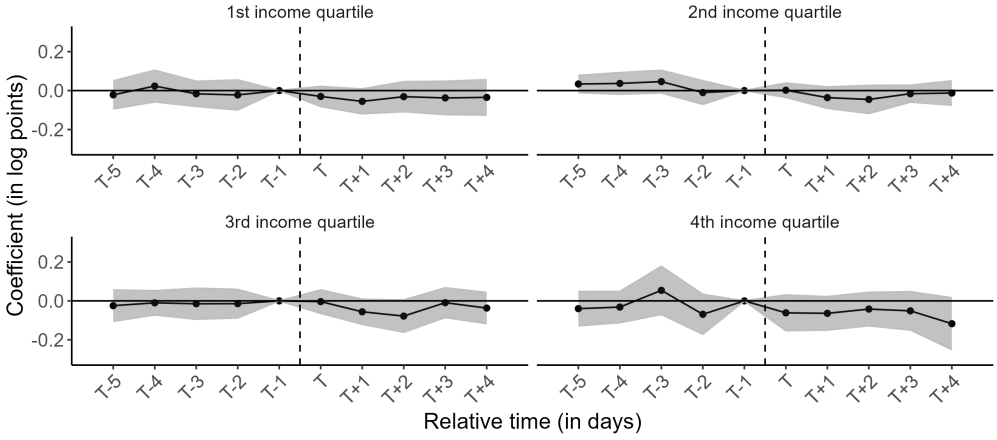
(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.25 – Main results from event-study design with residualisation specification without interaction

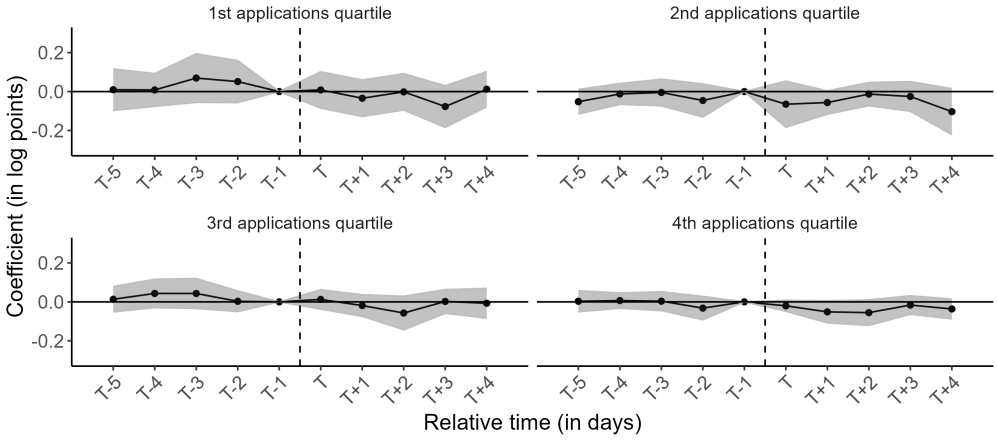
Notes: Panel 2.25a shows the results of the event-study regression of negative articles on the number of Universal Credit applications. The time series is first residualised by day-of-week fixed-effects interacted with a degree five polynomial in time (in days) and then restricted to 10-day windows, with only one-sided overlap between events allowed. Panels 2.25b and 2.25c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.



(a) Main regression



(b) Heterogeneity by regional quartile of income



(c) Heterogeneity by regional quartile of number of Universal Credit applications

Figure 2.26 – Main results from event-study design when dropping events on Saturdays.

Notes: Panel 2.26a shows the results of the event-study regression of negative articles on the number of Universal Credit applications when dropping events that occur on Saturdays. The time series is first residualised by day-of-week, month and year fixed-effects and then restricted to 10-day windows, with only one-sided overlap between events allowed. Panels 2.26b and 2.26c show the same approach when restricting to quartiles of regional income and regional level of Universal Credit applications respectively. All errors are clustered at the event level.

Chapter 3

Fixed-term contracts and wages: Rent-sharing and compensating differentials

1 Introduction

A central topic in European labor market policy debates in the past three decades has been about determining the appropriate level of employment protection legislation, with a particular focus on the regulation of fixed-term contracts (FTCs) and open-ended contracts (OECs). While academic discussions have primarily revolved around employment and labor market flows, as well as the productivity effects of FTCs, less attention has been given to their impact on the wage structure.[1]

The theoretical effect of FTCs on wages is ambiguous. Some theories suggest that FTCs could lower wages due to reduced bargaining power, while others argue that workers would need to be compensated for the decreased job security, resulting in compensating wage differentials. Understanding the mechanisms that determine wage formation in the presence of FTCs is crucial for understanding the full picture of their role in the labor market and the potential effects of

[1] Interestingly, this issue has also arisen in a recent German labor court case, where it was ruled that equal tasks should be remunerated equally, regardless of the contract type (see for instance https://www.spiegel.de/karriere/bundesarbeitsgericht-minijobber-muessen-den-gleichen-stundenlohn-bekommen-a-88f26250-208a-44 in German).

reforms.

To inform this debate, this paper presents a simple monopsonistic model of segmented labor markets categorized by contract type, augmented by the notion of compensating wage differentials for job security. It predicts several novel facts which I then document using comprehensive administrative data from France, which sets this study apart from much of the existing literature that relies on aggregated, cross-sectional data.

By introducing a trade-off between the compensating wage differentials due to the valuation of job security and differential rent-sharing between FTCs and OECs, my model makes four predictions. The first two predictions concern the within-firm wage gap between OECs and FTCs. First, it predicts that low productivity firms pay higher wages to FTCs than OECs, while the opposite is true for high productivity firms. This stems from the fact the compensating wage differential is constant, but rent-sharing increases with productivity. Low-productivity firms would then pay low wages to both OECs and FTCs, but they have to compensate FTCs for the lower job security, leading to a firm wage premium for FTCs in low productivity firms. The second prediction is that the overall sign of the FTC-OEC wage is ambiguous. This is a direct corollary of the first prediction as it depends on the productivity distribution of firms in the economy.

The next two predictions concern differential rent-sharing. The theory predicts firms share rents less with workers on FTCs than workers on OECs. This has been discussed before in the literature, but generally in the context of bargaining. Here it comes from a purely monopsonistic mechanism. Additionally, it also predicts that the degree of rent-sharing varies with the corresponding reduced-form firm labor supply elasticities.

I begin my empirical analysis by examining the aggregate wage gap in hiring wages. I focus on hiring wages to eschew the additional complexities that dynamics bring. I compute the overall gap in hiring wages between FTCs and OECs. After controlling for age, gender, occupation, industry and firm fixed effects, I find that the difference in hiring wages between FTCs and OECs is precisely estimated to be zero. This suggests that theories predicting both positive and negative wage gaps are necessary to understand the effect of FTCs on wages.[2] To further differentiate between these competing theories, I present several novel empirical findings.

[2] I exclude the possibility that this due simply to the mandate from the law to pay FTCs and OECs equally.

Given that my model emphasizes the firm dimension, I then analyze the behavior of the FTC-OEC wage gap at the firm level. I show wages in OECs and FTCs vary with value added per worker with different slopes and in the way predicted by the model, generating a wage premium for FTCs in low productivity firms and wage penalty for high productivity firms. To illustrate the compensating wage differential mechanism further, I show that wages in FTCs decrease with average FTC length in firms.

I then analyze the differential rent-sharing mechanism further. Using the approach pioneered by Card et al. (2016) for studying the gender gap, and recently applied to the outsourcing literature by Drenik et al. (2022), I compare the firm wage premia for FTCs and OECs separately, which are identified from worker moves(Abowd et al., 1999). This provides a measure of the degree to which high-wage firms for FTCs are also high-wage firms for OECs.

In line with the hypothesis of reduced rent-sharing among workers on FTCs, I find that, on average, firms offering a 10% wage premium to workers on OECs only provide a 5.1% premium to workers on FTCs during the period from 2010 to 2014. This percentage is very similar to the 4.9% premium observed by Drenik et al. (2022) in the outsourcing context in Argentina. These findings align with the predictions of the monopsonistic model regarding differential rent-sharing.

Finally, I explore the last prediction on the variation of rent-sharing by using variation in hiring concentration by contract type in local labor markets. First, I show that FTC markets are more concentrated that OEC markets. Second, I show that for OECs the wage-value-added slope (and hence rent-sharing) decreases as OEC concentration increases as predicted by the model. Finally, I show that FTC concentration seems to affect the wage-value-added slope for FTCs very little, which suggests that concentration is a less important source of market power for FTCs than OECs.

This paper contributes to the literature on FTCs by providing a monopsonistic point of view which emphasizes the role of the firm and differential rent-sharing. I argue that FTCs play a crucial role when trying to understand labor market power in dual labor markets. Additionally, I emphasize the importance of compensating wage differentials and the valuation of job security when studying FTCs. Finally, I document novel empirical facts in accordance with the previous considerations. The goal is that these motivate deeper investigations, both theoretical and empirical, into the effects of FTCs on the wage structure.

The rest of the paper is structured as follows. In the following section, I give a more thorough literature review which focuses on the wage effects of FTCs. Next, I present the model, flesh out its predictions and discuss its limitations. I then present the institutional details and the data. The last three sections are empirical. I start by analyzing the average FTC-OEC wage gap. I then study firm-level heterogeneity, using both a binscatter and a firm fixed-effect approach. And finally, I explore local labor market heterogeneity, looking at the effect of concentration on rent-sharing. The final section concludes.

2 Literature review

In this section, I provide a literature review focusing on the wage effects of FTCs. This is a fairly stringent restriction, as the literature has focused surprisingly little on wage formation with FTCs. This is in stark contrast with the literature on outsourcing, where the wage effects of outsourcing are a main concern.

Theoretical literature Early on, Bentolila et al. (1994) suggested an "insider-outsider" view of dual labor markets, where both temporary and unemployed workers should be considered as outsiders. Part of their mechanism is the degree to which unions represent the interests of temporary workers. The prediction from this model is a wage penalty for FTCs.

The early literature also looked at the learning component of FTCs, namely that they allow firms to learn the productivity of the worker. Blanchard and Landier (2002) formalize the latter by embedding it into a search-and-match model. They do not tease out any conclusions on the wage. Later literature partially dismissed this, arguing that in most countries OECs have trial periods with no firing costs. This objection rings particularly true for repeated and very short FTC use.

FTCs have also been studied in an efficiency wage perspective in Güell (2000); Güell and Rodríguez (2010), where incentives in OECs are provided by higher wages, whereas the renewal rate plays that role for FTCs. This also predicts a wage penalty for FTCs.

The most explored mechanism is that firms' use of FTCs could be explained by the fact that they face temporary production opportunities. Worker-firm matches in FTCs then have a lower average surplus which implies lower wages. Additionally, because terminating temporary

contracts before their date of termination is very costly, there can be situations where employers pay positive wages to unproductive temporary workers. This reduces their entry wage[3]. This has been modeled in Cahuc et al. (2016, 2020); Daruich et al. (2023), but the focus of all of these papers is the choice between FTC and OEC and the effect on employment and labor. Again, the prediction is a wage penalty for FTCs. Daruich et al. (2023) extend this model to give differential bargaining power to workers on FTCs and OECs.

On the other hand, Rion (2021) predicts that the wages hiring wages of OECs are lower than those of FTCs (but the continuing wages of OECs are the highest). Recently, Créchet (2023) has explored the role of risk-sharing between firms and workers with a trade-off between insurance and the flexibility of separation. Franceschin (2023) models a situation where excessive on-the-job search can be mitigated by the use of FTCs and OECs. One constant across most of these paper is that wages and in particular predictions on wages play a very minor role, and are rarely stated, making them hard to tease out.

Finally, a dimension which has not been explored in the context of FTCs and OECs is the role of job security and therefore compensating wage differentials. Jarosch (2023) presents a wage posting model with job security and shows that it can play an important role in understanding unemployment dynamics. Pessoa De Araujo (2017) looks at the role of job security in wage inequality. More generally, there has been a recent literature looking and documenting the importance of compensating wage differentials (Sorkin, 2018; Lavetti and Schmutte, 2018).

Empirical literature Over the past 20 years, several papers have looked at the wage gap between FTCs and OECs using different methodologies and data (see Table 1 in Albanese and Gallo (2020)). All of these papers find a consistent wage penalty for FTCs, even when looking at heterogeneity along different variables (with very few exceptions). An important potential issue with all of these studies is the use of survey data, which is limited by small sample sizes, potential misreporting and also limited control variables (for instance detailed occupation information or firm identifiers).

Recently, there have been a few papers that find opposite results when using more exhaustive data. Albanese and Gallo (2020) found that when using administrative data, restricting to hiring wages and using a stringent matching methodology based on extensive labor market history,

[3]As far as I understand the model, the prediction concerns the full wage rather than the hourly wage, which is of less interest since it compares wages of people over different employment lengths.

there is a significant and big wage premium for FTCs across the entire wage distribution in Italy. Using a structural approach, Lagrosa (2022) also found evidence for wage premia for FTCs in Spain. On the other hand, Daruich et al. (2023) find that a reform that lifted constraints on the employment of FTCs allowed firms to lower labor costs by decreasing earnings of young workers. The main mechanism they put forward is a differential rent-sharing effect.

3 Theoretical discussion

In this section, I outline a simple static model to provide a framework for thinking about the empirical results. I chose this model to emphasize several dimensions that are under-theorized in the FTC literature: the role of non-atomistic firms, a segmentation between the OEC and FTC markets and the role of a job security compensating differential.

3.1 The model

In this section, I describe the model. It is a straightforward combination of Cardoso et al. (2018) and Dube et al. (2022).

Worker preferences The main novelty for the FTC literature is that I assume that workers value job security. I introduce this valuation directly into the worker utility, as in Bonhomme and Jolivet (2009), rather than dynamically as in Jarosch (2023). The main reason is that it allows me to keep the model simple and static. I therefore follow Dube et al. (2022) in writing the indirect utility of a worker as:

$$V(w, \delta) \equiv w + \delta \tag{3.1}$$

The parameter δ represents the job security valuation. To keep the model as simple as possible (and easily solvable analytically), I have taken wage and amenity to be perfect substitutes. [4]

For simplicity, I take the δs as fixed, so that:

$$V^{\text{FTC}}(w) \equiv V(w, 0) = w$$

$$V^{\text{OEC}}(w) \equiv V(w, \delta)$$

[4]This could be relaxed by introducing a CES specification, but it precludes an analytical solution.

I assume that the labor market is segmented, namely that some workers only work in OECs and some only in FTCs. I discuss this assumption in more detail below. Workers then face an idiosyncratic taste shock z_j^{FTC} or z_j^{OEC} of working at firm j. The corresponding utilities are therefore given by:

$$u_j^{\text{FTC}} = z_j^{\text{FTC}} \left(V^{\text{FTC}}(w) - V_0^{\text{FTC}} \right), \qquad u_j^{\text{OEC}} = z_j^{\text{OEC}} \left(V^{\text{OEC}}(w) - V_0^{\text{OEC}} \right) \tag{3.2}$$

where V_0^{C} are the relevant outside options. I assume that the z^C are independent draws from Frechet distribution with shape parameter η^C, which implies from standard results in discrete choice theory, that the probability for a given worker of working for firm j is given by the expressions:

$$\begin{aligned} p_j^{\text{FTC}} &= \frac{\left(w_j - V_0^{\text{FTC}}\right)^{\eta^{\text{FTC}}}}{\sum_{i \neq j} \left(w_i - V_0^{\text{FTC}}\right)^{\eta^{\text{FTC}}}} \approx \lambda^{\text{FTC}} \left(w_j - V_0^{\text{FTC}}\right)^{\eta^{\text{FTC}}} \\ p_j^{\text{OEC}} &= \frac{\left(w_j - V_0^{\text{OEC}}\right)^{\eta^{\text{OEC}}}}{\sum_{i \neq j} \left(w_i - V_0^{\text{FTC}}\right)^{\eta^{\text{OEC}}}} \approx \lambda^{\text{OEC}} \left(w + \delta - V_0^{\text{OEC}}\right)^{\eta^{\text{OEC}}} \end{aligned}$$

where $\lambda^C = \frac{1}{\sum_{i \neq j} \left(w_i - V_0^{\text{C}}\right)^{\eta^{\text{C}}}}$. Under the standard assumption that $J \gg 1$, which abstracts from strategic interactions (as in for instance Cardoso et al. (2018)), I can consider the λ^Cs to be constant.

This analysis implies that firm j faces a *firm-specific* labor supply for each contract type given by:

$$\begin{aligned} L_j^{\text{FTC}}(w) &= N^{\text{FTC}} \cdot p_j^{\text{FTC}} = \left(N^{\text{FTC}} \lambda^{\text{FTC}}\right) \left(w - V_0^{\text{FTC}}\right)^{\eta^{\text{FTC}}}, \\ L_j^{\text{OEC}}(w) &= N^{\text{OEC}} \cdot p_j^{\text{OEC}} = \left(N^{\text{OEC}} \lambda^{\text{OEC}}\right) \left((w + \delta) - V_0^{\text{OEC}}\right)^{\eta^{\text{OEC}}} \end{aligned}$$

where N^C are the total amount of workers in contract C.

Remark. The same result could have been derived as in Cardoso et al. (2018) with a indirect utility of $\log(w + \delta - V_0) + e_i$ with e_i type-II extreme.

Firm problem I take the simplest possible case for the firm side: The firm faces a fixed output price P_j^0 and has a linear production function

$$Y_j = T_j f\left(L_j^{\text{FTC}}, L_j^{\text{OEC}}\right) = T_j\left((1-\theta)L_j^{\text{FTC}} + \theta L_j^{\text{OEC}}\right)$$

where T_j is a general TFP level of the firm and f is the production function. Additionally, this implies that workers on FTC and OEC are perfect substitutes and their relative productivity is governed by θ. I take $\theta \in [0.5, 1]$, which corresponds to the natural hypothesis that OECs are *more productive* than FTCs. This is a recurrent theme in the literature on FTCs . Finally, notice that under these assumptions, the value added by standardized unit of labor is given by $v_j \equiv \frac{P_j^0 Y_j}{f\left(L_j^{\text{FTC}}, L_j^{\text{OEC}}\right)} = P_j^0 T_j$.

The firm j problem is then given by:

$$\max_{\left(w_j^{\text{FTC}}, w_j^{\text{OEC}}\right)} v_j\left((1-\theta)L_j^{\text{FTC}} + \theta L_j^{\text{OEC}}\right) - \left(L_j^{\text{FTC}} \cdot w_j^{\text{FTC}} + L_j^{\text{OEC}} \cdot w_j^{\text{OEC}}\right)$$

Finally, continuing in the spirit of making this model the simplest possible, I follow Cardoso et al. (2018) in taking the outside options to be given by

$$V_0^{\text{FTC}} = (1-\theta)\cdot b, \qquad V_0^{\text{OEC}} = \theta \cdot b$$

This corresponds to a situation where the outside options are given by an outside wage that could be earned in another competitive sector, with the same productivity differential indexed by θ.

3.2 Predictions

The FOC yield:

$$w_j^{\text{FTC}} = \frac{1}{1+\eta^{\text{FTC}}}(1-\theta)\cdot b + \left(\frac{\eta^{\text{FTC}}}{1+\eta^{\text{FTC}}}(1-\theta)\right)\cdot v_j, \quad \text{for } v_j > b$$
$$w_j^{\text{OEC}} = \frac{1}{1+\eta^{\text{OEC}}}(\theta\cdot b - \delta) + \left(\frac{\eta^{\text{OEC}}}{1+\eta^{\text{OEC}}}\theta\right)\cdot v_j \quad \text{for } v_j > b - \frac{\delta}{\theta}$$

This then yields the several predictions.

Prediction 1. *Depending on the magnitude of the job security amenity δ, there can be 3 cases:*

(i) If δ is small, FTCs are always paid less than OECs

(ii) If δ is large, OECs are always (in a reasonable range of v) paid less FTCs

(iii) For intermediate values of δ, low productivity firms pay higher wages to FTCs than OECs. The opposite is true for high productivity firms. Namely, there exists $\tilde{v}$ such that:

$$\textit{If} \quad v < (>)\tilde{v} \quad \textit{then} \quad w^{\mathrm{FTC}} > (<) w^{\mathrm{OEC}}$$

Prediction 2. *The sign of the average wage gap is therefore undetermined and can be zero.*

These first two predictions concern the within-firm wage gap between FTCs and OECs. The next two predictions are about differential rent-sharing.

Prediction 3. *There is differential rent-sharing within firm between OECs and FTCs. Namely,*

$$0 < \frac{\mathrm{d}w^{\mathrm{FTC}}}{\mathrm{d}v} < \frac{\mathrm{d}w^{\mathrm{OEC}}}{\mathrm{d}v} \qquad \textit{if} \qquad \eta^{\mathrm{OEC}} > \eta^{\mathrm{FTC}}$$

Prediction 4. *The value-added slope contract type C increases when η^C increases. In other words:*

$$\frac{\mathrm{d}}{\mathrm{d}\eta^C}\left(\frac{\mathrm{d}w^C}{\mathrm{d}v}\right) > 0$$

The last prediction states variation in η should yield variation in differential rent-sharing as represented by the slope between wages and value-added.

3.3 Interpretation of the model

The economic mechanism behind the model can be summarized as follows. The upward sloping labor supply curves from monopsony imply positive correlation between productivity and wages. On the other hand, OECs come with a constant utility premium because of the job security. It is therefore possible that low-productivity firms that pay low wages have to pay higher wages to individuals on FTCs.

The different firm specific labor supply elasticities play a crucial role in this model. An important question is then the source of the different labor supply elasticities. As is standard

in the monopsony literature, in the model presented above they come from the shape of the preference distribution of individuals, in other words from preferences over non-pecuniary job characteristics. But ultimately, the predictions rely on the reduced-forms of the labor supply curves, and it is therefore worth exploring other sources for the differential labor supply elasticities. Standard mechanisms in the literature are search frictions and mobility costs. These would seem to be lower for FTCs and would therefore suggest lower market power for FTCs. Another source which is often discussed is training cost and job-specific human capital. This has also been discussed extensively in the FTC literature. Indeed, given their short-term nature, FTCs discourage training and investments in job-specific human capital. This would lower the market power of workers on FTCs. This mechanism underlies for instance the differential rent-sharing in Kline et al. (2019) between new and incumbent workers.

Bargaining models can generate very similar results to monopsony models. Lower bargaining power for FTCs is sometimes discussed as a mechanism. One reason which is mentioned in Daruich et al. (2023) is the possibility that FTC workers are underrepresented by unions and firm-level wage agreements (this is also discussed in Bentolila et al. (1994)). Another possibility for lower bargaining power is due to the higher precarity of individuals on FTCs.

Another interesting mechanism is the explicit segmentation of the labor market introduced by the presence of FTCs. In general, the hypothesis that the FTC and OEC market are at least partially segmented seems reasonable. Firms generally post different vacancies for FTCs or OECs. Additionally, the pools of relevant and targeted individuals is different. There is certainly mobility between the markets, in particular from the FTC market to the OEC market, but this is still consistent with some level of segmentation. It is conceivable that this allows for stronger price discrimination for firms. It also potentially creates norms around low-wage labor market sectors.

Finally, differences in outside options between individuals on FTCs and OECs could also play an important role. For instance, it is possible that for individuals on OECs the general outside option are other firms for job-to-job transitions, whereas for FTCs the outside option is unemployment benefits. Additionally, it is also possible that the (mis)perception of outside options is different between OECs and FTCs, which could have a significant effect as was shown recently by Jäger et al. (2023). These could also reinforce the segmentation between the two labor markets.

3.4 Discussion of the model

In this section, I discuss the different omissions and weaknesses of the model presented above.

Dynamic considerations The model I present is a static model. It allows me to make the basic predictions of interest. On the other hand, it is clear that dynamic considerations should play an important role for FTCs. One obvious indicator for this, is that one of the main debates in the FTC literature is whether FTC are low-wage traps or stepping stones towards better jobs, which is an inherently dynamic consideration. Recently, Jarosch (2023) argued that job security is crucial to understand unemployment and in particular the "unemployment scar", because it creates a job ladder with "slippery" bottom. Job security, in particular at the bottom of the wage distribution, is very related to FTCs.

Another interesting hypothesis is that the existence of FTCs creates a dual wage ladder: while on FTCs, workers experience essentially no wage growth, and essentially only step onto the wage ladder once they are on an OEC. In other words, FTCs create a labor market segment where wage growth is absent. This expected wage progression then for instance predicts a high number of wage cuts when switching from an FTC to an OEC (as is standard in the Postel–Vinay and Robin (2002) literature).

Another dimension which is missed in a static model is the role of business cycles in the utilization of FTCs. FTCs are generally the first workers to leave a firm in a recession and as such they are a way for firms to ensure flexibility against shocks. This dimension has been explored in the literature, which in particular finds that firms in more volatile industries use more FTCs (see for instance Caggese and Cuñat (2008)).

It is clear from this discussion that an interesting avenue for future research would be to microfound the model above by for instance embedding it in search model. It would be of particular interest to endogenize δ. An interesting option could be to allow for bargaining (or posting) on wages and contract length.

Institutional context Another significant omission in the model is any institutional context. From the point of view of effect on wages, there are three instutions which seem of particular interest.

First, the existence of the prime précarité and the exemptions to it introduce interesting

potential strategic considerations. The fact that certain sectors are exempt potentially introduces distortions between sectors but also has potential spillover effects across industries. The degree to which firms ''incidence" away the prime précarité is also an interesting question.

Second, the minimum wage and sectoral wage minimas will affect in particular the firms ability to ''incidence" away the prime précarité. For instance, for jobs at the minimum wage an FTC will necessarily be more expensive than an OEC, which suggests an additional incentive to consider for the use of OECs.

Finally, unemployment insurance will affect wage formation in a dual labor market. In particular, because unemployment insurance is an outside option which might be more salient for individuals on FTCs, than for individuals on OECs, for whom their current or equivalent jobs might be more salient. Additionally, rules such as partial unemployment insurance might increase the use of FTCs, both from a firm and a worker perspective.

4 Institutional details and data

4.1 Description of legislation

In France there are three primary types of employment contracts: open-ended contracts (OECs) or Contrat à durée indéterminée (CDI), fixed-term contracts (FTCs) or Contrat à durée déterminée (CDD), and interim (outsourcing) contracts.

The OEC is the standard contract and is the most common type of contract, accounting for around 85% of employment. These contracts have no specified end date, and employers cannot unilaterally terminate them without a valid reason. Termination of an OEC can only be justified by economic reasons, which must be substantiated and involve specific actions by the firm, or by personal reasons such as poor performance or misconduct. Consequently, firms face significant costs when terminating these contracts.

In contrast, FTCs have a predetermined end date, which is set when the contract is signed. The rules governing the termination of FTCs are as stringent as those for OECs, implying high termination costs before the designated end date. The law states that FTCs can only be used for specific purposes, such as replacing an absent employee, accommodating a temporary increase in activity, or facilitating seasonal employment. The duration of an FTC is limited by the relevant

Collective Bargaining Agreement (CBA)[5], which also specifies the number of contract renewals allowed. Generally, an FTC cannot be renewed without a valid reason, and there is a mandatory waiting period before re-employing someone on an FTC for the same position.

Furthermore, the law stipulates that wages for the same job should be identical between FTCs and OECs. If an FTC is not converted into an OEC, employers are required to pay employees an additional precarity bonus (prime précarité) equivalent to 10% of the total value of the contract. However, certain CBAs may lower the percentage of the prime précarité, and the bonus is not applicable if the employee declines the conversion to an OEC for the same job.

There is the additional case of CDD d'usage (CDD-U), which is a sectoral exemption to most of the rules mentioned above, in particular the payment of the prime précarité. This exemption was introduced to simplify FTC use in industries where FTCs are ''standard''. The sectors which are allowed to use CDD-U are fixed by a list, but the use and importance of the phenomenon is notoriously opaque and therefore hard to quantify (see for instance Marie and Jaouen (2015)).

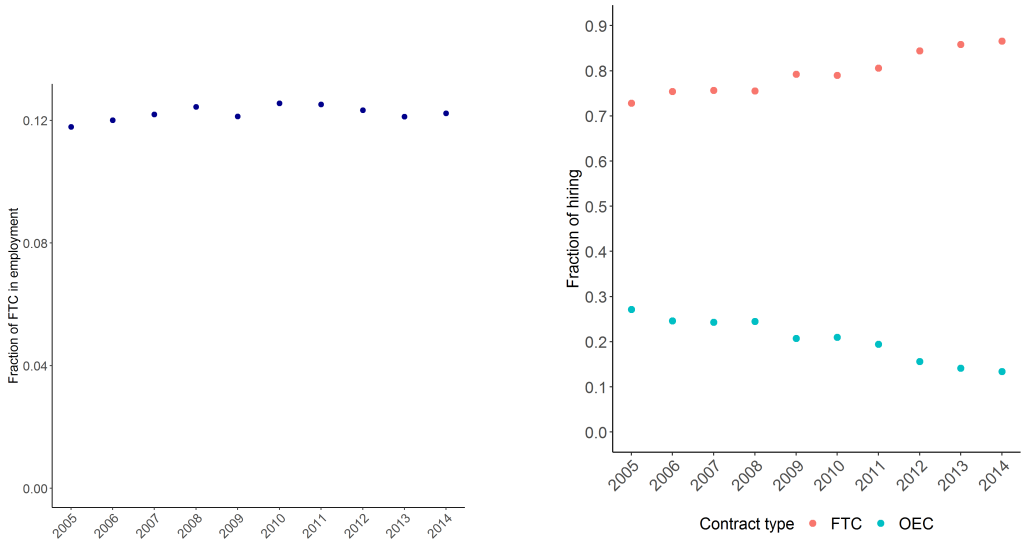
4.2 Descriptive facts

Although FTCs accounted for approximately 12% of employment (only among FTC and OEC workers ie excluding the small fraction in outsourcing and other contracts such as apprenticeships) in the period 2005 to 2014, their short-term nature results in them representing up to 90% of hires, as depicted in Figure 3.1. In terms of overall employment, this constitutes about five times the employment observed in the outsourcing sector, which represents around 2% of employment, highlighting the significant role of FTCs in the French labor market.

For both firms and individuals, FTCs cover diverse uses and circumstances. At the firm level, 3.2a shows that there is a lot of variation in the average FTC hiring rate between firms. Among all firms, 25% of firms never hire on FTCs while for bigger firms this drops to 5%. Among the latter these larger firms only around 20% of firms hire less than 25% on FTCs whereas around 40% hire at least 75% on FTCs. Figure 3.2b shows that average FTC length by firm is heterogeneous but more concentrated than use. Indeed, more than 50% of firms have an FTC length of less than 3 months[6]. An interesting feature of this figure are the kinks are the 3-month, 6-month and

[5]In general, an FTC has a maximal allowed length of 18 months, but this can vary depending on the stated reason for use of FTC, which allows in some cases for 24 months or restricts further to 9 months.

[6]These represent an upper bound as the graph uses contracted length rather than actual length.



(a) FTC employment share **(b)** FTC and OEC hiring shares

Figure 3.1 – FTC aggregate time series

Notes: Panel 3.1a shows fraction of employment in FTC among FTC and OEC. It is computed from the statistically matched DADS panel described in the data section 4.3 by taking the fraction of worker in an FTC (among OECs and FTCs) at the first of each month and then averaging over the year. Panel 3.1b shows the fractions of FTC and OEC hiring (among OECs and FTCs) computed from the DPAE data described in the data section 4.3.

12-month average lengths, suggesting a standardized use of FTC length[7].

Another important dimension of FTC employment in firms is the rate of conversion into OECs which, as was argued above, is important for the case where FTCs would be used for learning about workers or to provide incentives on FTCs. Aggregate computations based on hiring data[8] yield that over the years 2005 to 2014, only around 6% of matches (defined as a match between an individual and an establishment) which had at least one hire in FTC, also had a hire in OEC. In other words, only a small minority of FTCs are converted. At the firm level, conversion rate can be defined in two ways. In figure 3.2c, I show the FTC conversion rate defined as the fraction of FTCs hired in a year that will eventually (within the period) get converted into an OEC. It shows that the FTC conversion rate is concentrated at low values, but that there is again significant heterogeneity, with for instance around 25% having an FTC conversion of 20% or more. In the appendix, I show an alternative way of measuring conversion

[7]In particular, it generates some question around the interpretation that firms need specific temporary contracts. An interesting potential connection is that the rounding in FTC length is an indicator of market power as Dube et al. (2018) argue about the rounding of wages.

[8]The DPAE data described in the data section below.

rate: the OEC conversion rate defined as the fraction of OECs hired in a year that were previously employed in an FTC at the firm. While still concentrated at the lower end, it shows a significantly wider dispersion. An important conclusion of this descriptive paragraph, which is one of the themes in this paper, is that FTC use and practices are very heterogenous.

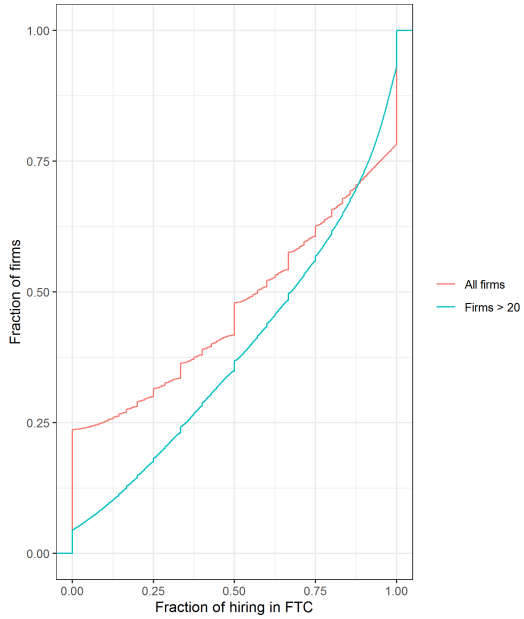
At the individual level, Figure 3.3a shows the variation of the fraction of FTC employment in 2-digit and 4-digit occupations. The occupation with the highest FTC use is Information, arts and entertainment professionals (code 35). Within the worker category (codes starting with 6) the FTC employment fraction varies from around 5% for Skilled industrial workers to around 25% for Agricultural workers. The graph illustrates the important role played by occupation in explaining FTC use and the fact that 2-digit occupation codes capture a good part but not all of that variation.

Figure 3.3b shows the variation of FTC employment by age. This fraction drops dramatically from nearly 70% at age 18 to around 10% at age 30. The fraction stays roughly constant until age 60 after which it increases again, which is could be due to early retirements or FTCs for seniors (CDD senior). This illustrates that age plays an important role, but between 30 and 60 the FTC employment rate of 10% still represents a sizable share. While comparisons across countries are complicated because of different definitions and legislations pertaining to FTCs, it does seem to be generally true across Europe that young people are particularly employed in the FTC (see for instance Matsaganis et al. (2014)). The reasons for this pattern are not immediatly clear. On the one hand, it could reflect preferences of young workers for more flexible contracts. More likely, is that firms are less inclined to hire young workers on OECs and young workers have lower bargaining power. Exploring this pattern more deeply is an interesting avenue for future research.

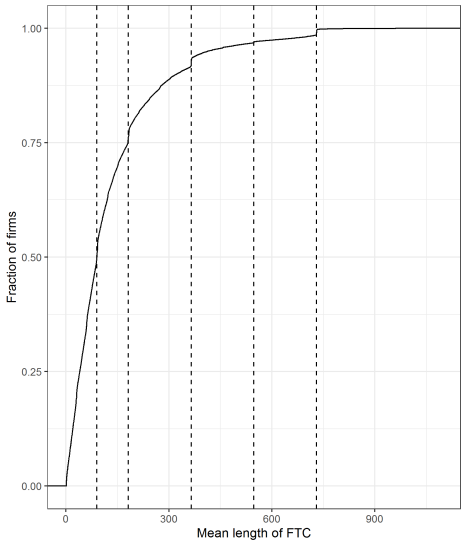
4.3 Data

The empirical results in this paper rely on multiple administrative data sources.

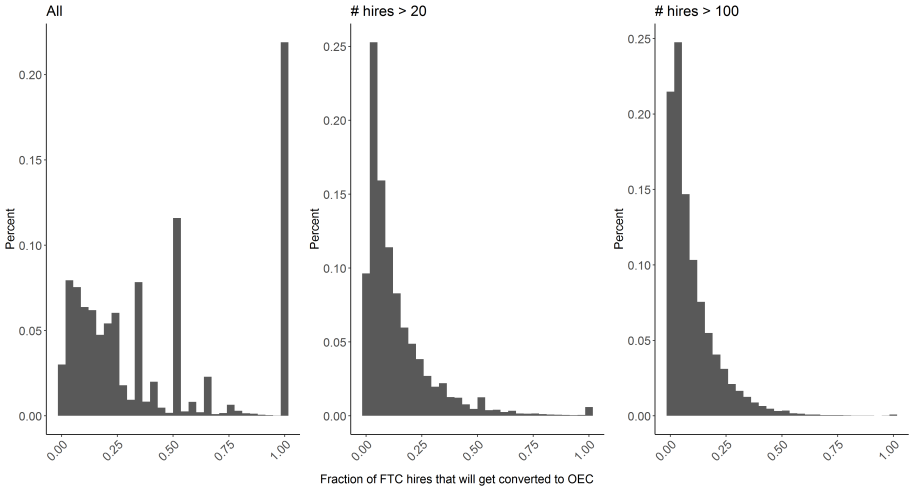
Matched employer-employee data The primary data source for this study is the employer tax records, specifically the "Déclaration Annuelle de Données Sociales" (DADS). To construct a comprehensive matched employer-employee panel spanning from 2005 to 2018, I use the statistical match approach developed by Babet et al. (2022). Their approach utilizes the 2-year



(a) Distribution of average FTC hiring share by firm



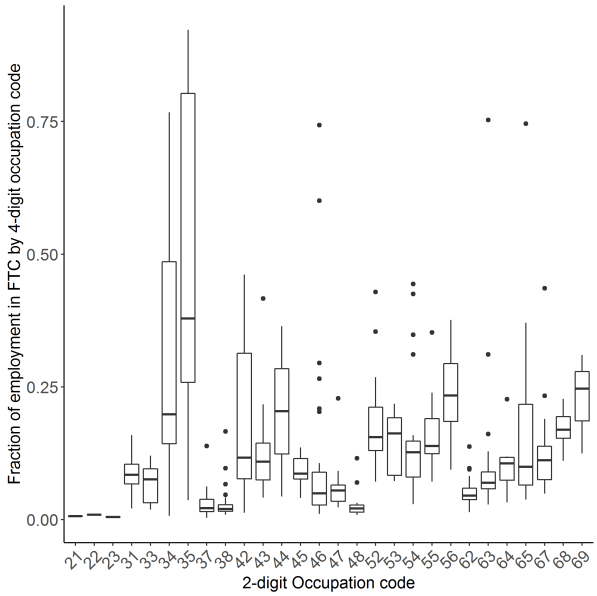
(b) Distribution of average FTC length in days share by firm



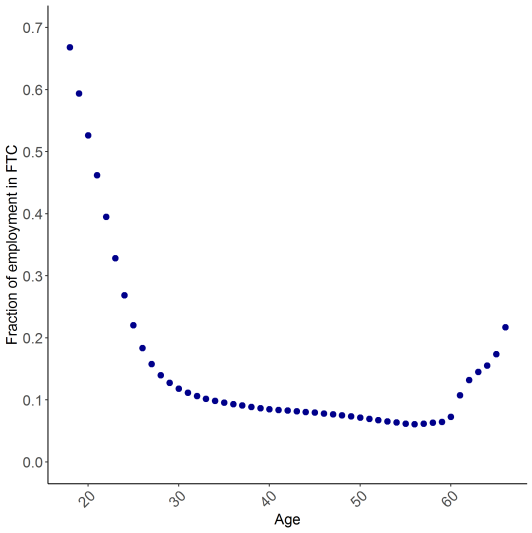
(c) Distribution of the fraction of FTCs hired in a year that will get converted into an OEC

Figure 3.2 – Distributions in different measures of FTC use at firm level.

Notes: Panel 3.2a shows the firm-level distribution of the average FTC hiring rate (among FTCs and OECs), averaged over the entire 2005-2014 period, for all firms (red line) and firms bigger than 20 (blue line). Panel 3.2b shows the firm-level distribution of average FTC length in days, averaged over the entire 2005-2014 period. The dashed lines mark 3-month, 6-month, 12-month, 18-month and 24-month contract lengths. Panel 3.2c shows the distribution of the FTC conversion rate defined as the fraction of FTCs hired in a year that will eventually (within the period) get converted into an OEC, among all firms, firms with more than 20 and firms with more than 100 hires. All the figures use the DPAE data described in section 4.3.



(a) Boxplot of FTC employment fraction by 2-digit occupation code, with 4-digit code variation.



(b) Fraction of employment in FTC by age

Figure 3.3 – Different measures of FTC use at the worker level.

Notes: Panel 3.3a shows a boxplot of average FTC employment by 2-digit occupation code. The variation within each 2-digit occupation code comes from the underlying 4-digit occupation codes. The occupation code description is given in table 3.6. The bar represents the mean, the hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 times the inter-quartile range from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 times the inter-quartile of the hinge. Data beyond the end of the whiskers are called outliers. Panel 3.3b shows fraction of FTC employment by age.

panel form of the DADS postes files, which allow very detailed matching of individuals across year-files[9]. In the period I am studying the match quality is at 98%. One particular issue with the statistical match is that for individuals that are out of the labor market for more than a year the match breaks down as it cannot connect them anymore and they will therefore be identified as separate individuals. This should only affect a small minority of individuals. Nevertheless, I will present some robustness results using the cross-sectional DADS postes data. I discuss why my preferred approach still uses the panel version of the data just below.

This dataset then provides detailed information on each worker in the firm: total annual wages, number of hours worked, start and end date for each job, contract type, occupation, as well as individual information such as gender, age and city of residence.

I use the measure of gross wages, adjusted for inflation to 2012 euros, and create a log hourly

[9]For more details on the procedure, see Appendix C in Babet et al. (2022)

wage variable by dividing by the number of hours. I also trim the wage measure at the 0.1% level on both sides. For most of my analysis, I focus on hiring wages. Hiring wages on OECs are defined straightforwardly as the hourly wage in the first year working for the firm.[10] For FTCs the definition is slightly more complicated due to the existence of the prime précarité. Indeed, for any FTC spell that covers more than a single year, the wages defined as for OECs would be too low as they would not include the prime précarité (because it would be included only in the corresponding observation in the year the corresponding FTC spell ends in). To address this, I define hiring wages on FTCs as the wages in the first continuous spell working for the firm. This is also the reason why my preferred approach uses the panel data.

I restrict the dataset to the private sector, individuals aged between 18 and 66 and mainland France. For the main analysis, I focus on the years 2010 to 2014. I start in 2010 because before 2009, it was not mandatory to report the 4-digit occupation, which is an important control in the following regressions. Indeed, the 2-digit occupation codes separate occupations into 29 categories, which can be fairly broad.[11] Additionally, I limit the data to CDD and CDI, specifically excluding all individuals on interim (outsourcing) contracts.

To capture relevant work experience, I create two experience variables. The first variable, experience in firm, measures the number of years since the worker was initially hired by the firm. It measures the worker's tenure with the specific employer. The second variable, experience in (2-digit) occupation, measures the number of years since the worker was first hired in their current occupation. This variable helps to account for the worker's accumulated experience in their specific field of work.[12]

The classification of observations into OEC and FTC will contain some measurement error. Firstly, the contract type is self-declared in the DADS. More importantly, the DADS collapses repeat spells of an individual in the same year at the same establishment, and in particular during this collapse, if there is at least one spell in OEC, then that observation is classified as an OEC. In other words, individuals that are converted from an FTC to an OEC in the same year will be classified as only an OEC. I believe the effect of this on the results that follow should be limited because of the low rate of conversion of FTCs into OECs highlighted in the previous

[10] This does not mean that the wages cover a year worth of wages in the firm. Indeed, for somebody hired in December, the wage comes from that one month of employment.

[11] The code 37 for instance covers all jobs with management responsibilities in administrative and commercial matters, which includes lawyers, accountants, PR managers, HR managers and more.

[12] Both these experience measures have the defect that interruptions are not considered and experience continues accumulating. This is only a problem for the experience in occupation variable, since returns to firms are fairly rare.

section.

There are two complications in the measurement of wages, which I cannot deal with directly. First, for FTCs, wages include the prime précarité. In other words, the prime précarité is not measured separately. The correct wage concept includes the prime précarité. However, as detailed above, there are selected instances when the prime précarité does not have to be paid, which could introduce omitted variable bias. The first case is when an FTC is converted to an OEC, as discussed above. Given the very low levels of conversions, this should not be too much of a problem. A more significant problem comes from the use of CDD-U (described in section 4.1) which are exempt from paying the prime précarité, but are hard to identify. I provide a robustness check for this concern in section 5.3. The second concern stems from the measurement error in hours. As was already mentioned, there is partly some measurement error in hours stemming from missing hours which are imputed by the data producer. But there is also an additional concern from the fact that the hours measure in the DADS contains paid holiday hours. The rules for FTCs and OECs are slightly different (and complicated), and leave not taken can be compensated. This could introduce distortions.

Contract declaration data My second data source is the contract declaration data, ''Déclaration Préalable à l'Embauche''(DPAE), which has to be reported whenever a new contract is signed. This dataset provides a comprehensive individual-level panel that includes information about all newly signed contracts, including the starting date and type of the contract, as well as the establishment where the contract was signed. For FTCs, the DPAE also includes the specified end date of the contract. However, it does not provide any wage information. It is unfortunately not possible to link DPAE and DADS data. In particular, comparing results between the DADS and the DPAE of similar measures shows that there are some differences between these datasets. In part this is due to the fact that for instance within a year and an establishment, multiple employment spells are merged in the DADS, but also less clear differences. For the relevant results below, I run a robustness check and show that the results are unaffected.

Until the year 2009, there is an issue in part of the data with missing end dates for FTCs as well as issues with grouping of individual identifiers. In the main analysis, I drop these observations. I conducted some robustness checks (not shown) and the differences are small.

Financial data My third data source is the near-universe of annual tax records of French firms ("Fichier Complet Unifié de Suse", FICUS, and "Fichier Approché des Résultats d'Ésane", FARE) that report balance sheet and income statement information. I use this dataset for firm-level measures of revenue and value-added. The main variable I use in the paper is the value added net of taxes, which I trim at the 0.1% level. The value added per worker variable is then defined by dividing value added by the effective average firm size over the year. This measure is effective because it is a legally defined notion of size which assigns certain weights to temporary workers.

5 Average FTC-OEC wage gap

As was argued above, the effect of FTCs on wages is theoretically ambiguous. Motivated by the predictions of the model, I present several empirical facts. In this section, I focus on hiring wages because as discussed above, wage dynamics significantly complicate the appropriate comparisons and harder to capture in the aggregate approach of this section. In the appendix section A.3, I provide a very brief look at wage dynamics.

5.1 Methodology

To get a first aggregate characterization of the FTC-OEC wage gap, I regress log hourly wages on an indicator for being on an FTC, controlling for progressively more variables to understand what covariates play an important role. The regression reads:

$$\ln hw_{it} = \beta \cdot \mathrm{FTC}_{it} + \tau_t + \left(X_{it} + \psi_{J(it)}\right) + \varepsilon_{it}$$

where FTC_{it} is an indicator for the employment spell being in an FTC, τ_t are year fixed-effects and $\psi_{J(it)}$ is an establishment fixed-effect where $J(it)$ represents the establishment of worker i and time t. The covariates in X play an important role in this analysis, as the goal is to obtain a wage gap for individuals that are as comparable as possible. I discuss the specific choice of covariates below. The regression above is used to generate some descriptive facts about the FTC-OEC wage gap. It should not be interpreted causally.

A first potential issue is that the parametric choice for the regression specification will only imperfectly control for the effect of the chosen observables. There is also a significant potential

for omitted variable bias as our controls are unlikely to capture unobserved components of match quality or potentially relevant features of the labor market history of the worker. Finally, the decision about whether to add individual fixed-effects merits a separate discussion. On the one hand, they are a powerful tool to address time-invariant confounders, which are hard to capture using controls. But importantly, they change the identification of the wage gap as it is then identified from comparison of switches between FTCs and OECs *within* individuals. These are then affected by the dynamic considerations I have already mentioned, and therefore answer a different question than what we are interested in this section. In particular, using individual fixed effects necessarily implies that individuals are no longer compared at the same age and experience.

5.2 Results

The results from the regression are shown in Table 3.1.

Table 3.1 – FTC-OEC hiring wage gap regression over the period 2010 to 2014

	Log hourly wage						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
FTC	-0.1311***	-0.1215***	-0.0673***	-0.0701***	-0.0037***	0.0058***	-0.0006
	(0.0022)	(0.0022)	(0.0021)	(0.0014)	(0.0007)	(0.0006)	(0.0006)
Observations	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570
R^2	0.02689	0.04358	0.12355	0.27117	0.54488	0.55270	0.67677
Within R^2	0.02682	0.02333	0.00747	0.00879	3.52×10^{-5}	8.49×10^{-5}	6.92×10^{-7}
Year fixed effects	✓	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓	✓
Industry fixed effects				✓	✓	✓	✓
Occupation fixed effects					✓	✓	✓
Exp. in occ. fixed effects						✓	✓
Establishment fixed effects							✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%, and FTC wages are aggregated over the FTC spell. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

The gap without any controls stands at -13%. Controlling for age halves that gap. This captures a general experience effect, in part consistent with the FTC use by age highlighted above. Interestingly, industry does not alter the gap significantly. The gap reduces to almost zero once occupation is controlled for. This is not surprising as FTCs are more used in lower-paid occupations. The gap is a very precisely estimated zero when restricting to comparison within firm by controlling for firm fixed-effects. In the appendix, I show that this results does not change when using less fine 2-digit occupation codes (Table 3.7) and looking at the earlier period between 2005 and 2009 (Table 3.8).

5.3 Discussion and robustness

The main conclusion from this table is that several of the different mechanisms mentioned above have to play a role since they are compensating each other on average. In particular, this confirms the possibility raised in Prediction 2. As such, this average zero actually conceals more than it reveals and a systematic heterogeneity analysis is necessary to gain more insights into the mechanisms

Expectation from the law The intent of the existing law is a priori fairly clear. It requires that two individuals with identical ''qualifications" and identical job should be paid the same independently of contract type. Importantly, as was confirmed by the courts[13], this *does not* include the prime précarité. In other words, according to the law, when comparing individuals in identical jobs and same qualifications, the individual in an FTC should have an hourly wage premium of 10%. This decreases on average once the exceptions to the prime précarité are considered such as for instance the CDD-U, where the prime précarité does not have to be paid (see Section 4.1 for more details).

Even before considering the results, it is unclear how this should play out in practice. First, the definition of identical job and qualification is somewhat vague. Second, it is unclear how well-informed workers are about their rights to the prime précarité. More importantly, the enforcement is in general up to workers as they can make their claim in front of the courts. But this is costly and complicated for workers.

[13] see for instance `https://www2.liaisons-sociales.fr/210-10-quelles-sont-les-regles-a-respecter-e` French, accessed in August 2023)

The aggregate zero from Table 3.1 can therefore have different factors influencing it. It could be consistent with the prime précarité being paid rarely, resulting in a zero gap. On the other hand, it could also be consistent with the prime précarité premium, being (illegaly) ''incidenced'' away by firms, through lower wages. While these considerations certainly play a role, I argue below that the mechanisms laid out in the theory play a more important role because of the heterogeneity patterns I observe.

Panel versus cross-sectional data To alleviate concerns about the statistical match panel data (such as the quality of the match or the panel interruption due to prolonged inactivity as discussed in section 4.3), Table 3.9 presents the same regression as 3.1 using the cross-sectional DADS postes data. As discussed previously, this implies that FTC wages are not aggregated over the spell, resulting in potential under-estimation because the prime précarité is missed. To ''control'' for this effect, Table 3.10 shows the same regression using the statistical match data, but without aggregating FTC wages over the spell. The regression with all fixed-effects yields a large and significant negative gap of -1.86%, but it is identical to the estimate using the panel data in Table 3.10. This therefore suggests that using the cross-sectional data rather than the panel makes it little difference *except* for the error related to not aggregating FTC wages, which appears significant. Indeed, the significant decrease from 0% to around -2% is consistent, at least in sign, with a significant underestimation of FTC wages when not aggregating over the spell, as was discussed in section 4.3. This shows that the use of the panel data should be preferred.

Age restrictions Figures 3.3b showed that age plays an important role for employment in FTCs, representing a large share of employment for individuals under 30, stabilizing thereafter and increasing again after age 60. I therefore present the result of the same regression as in Table 3.1, but restricting to individuals aged between 30 and 60 in Table 3.11. The FTC-OEC gap with all controls is very slightly positive at 0.6%, but the qualitative conclusion of a zero average wage gap is unaltered. This is not to say that understanding the exact role of age in the use of FTCs is not in part at least related to other mechanisms and worth exploring. This is left for future research.

Dealing with CDD-U As noted earlier, accounting for CDD-U is complicated because the legislation itself is extremely unclear. The criteria don't align with industries or jobs and there is

suspected wide misuse of CDD-U (see for instance Marie and Jaouen (2015)). I therefore use the very rough classification by industries developed in Marie and Jaouen (2015)[14] to separate the data into industries that are concerned by CDD-U and sectors that are not. This classification is very rough as many of the industries should only be very partially affected.

Table 3.12 shows that the results of the regression with all controls for these two sets of industries are essentially identical at around 0%[15]. If anything, the signs go in the opposite direction of what would be expected.

Effect of measurement error I discussed the measurement error due to the misclassification of FTCs as OECs within a single firm spell and argued that it should be minimal given the low numbers of within-firm conversions. The sign of the effect of this misclassification depends on the sign of the OEC-FTC wage gap, but it will tend to decrease the gap. While this pushes the aggregate wage gap towards zero, it also suggests that the heterogeneity that we document below is if anything understated.

6 Firm-level heterogeneity

Prediction 1 of the toy model above emphasized role of firms in understanding wages in the presence of FTCs and in particular the within-firm dimension of the wage gap. In this section, I therefore study the FTC-OEC gap at the firm level.

6.1 Binscatter approach

I create a firm level dataset by collapsing the data at the firm and year level. The data is then merged with the financial data from the FARE-FICUS dataset. This procedure involves selection on firms, as for example I only keep firm-year observations that hired both FTCs and OECs, but also the financial data is slightly more imperfect for smaller firms. The selection is shown in the

[14]They use the A38 classification into industries and consider A (agriculture, sylviculture et pêche), C (industrie manufacturière), F (construction), H (transport et entreposage), I (hébergement et restauration), J (information et communication), K (activités financières et d'assurance), M (activités spécialisées, scientifiques et techniques), N (activités de services administratifs et de soutien), P (enseignement), Q (santé et action sociale), R (arts, spectacles et activités récréatives) to be the industries that are potentially allowed the use of CDD-U

[15]The estimate on the full sample is slightly different than the one in Table 3.1 because the regression is computed on the sample used in the heterogeneity sample after merging with financial data, which results in some data loss. Importantly, the estimates are nearly identical.

appendix Table 3.13, which shows that it selects bigger firms and the average wage is slightly lower.

Correlation with valued added per worker In figure 3.4, I explore Predictions 1 and 3, by looking at the correlation between wages in FTCs and OECs and value added per worker. The figure shows binscatters of average firm level log hourly wages in FTCs and OECs against log value added per worker. The first panel presents raw data. In the second, wages are residualized on year, gender, age, occupation, experience in occupation and industry. In the third panel, log value added per worker is additionally residualized on the same variables. These figures are produced using a dataset where I have trimmed the bottom 10% of the value added distribution. I discuss this choice in more detail below.

All three panels exhibit a behavior strikingly in line with the predictions of the simple model above. The slope of FTC wages with valued added per worker is significantly lower than for OEC wages, as predicted by Prediction 3. Once wages are residualized, Figure 3.4 shows that low valued added per worker firms pay higher wages to FTC workers whereas the opposite holds true for high valued added per worker firms. The additional residualization of value added per worker significantly increases the FTC slope, which suggests that especially for FTCs wages and value added correlated to industry and occupation.

The slopes in Figure 3.4 can be quantified to give a very rough estimate of firm-level labor supply elasticities. The appendix Table 3.14 reports an elasticity of around 8.5% for OECs and 5.4% for FTCs (when both wages and value added are residualized). Interestingly, the OEC value is on the higher end of the estimates in the literature, whereas the value for FTCs exactly onto the commonly cited value of value (see for instance Manning (2021)).

In the appendix Figure 3.9, I show the same figure but for individuals aged 30 and above. The conclusions are the same but the magnitudes are different. The gaps at both ends are smaller and the elasticities are 9.8% for OECs and 7.8% for FTCs[16] This suggests that the labor market power of firms decreases significantly when eliminating younger workers.

In the appendix Figure 3.10, I show the same figure but without trimming the bottom 10% of the value added distribution. The figures show that the bottom 10% exhibit an opposite behavior to the rest. A quick analysis did not reveal any obvious covariate that predicted this behavior. A more detailed analysis of this fact would be interesting, but is left for further research.

[16]Regression table not shown in this document.

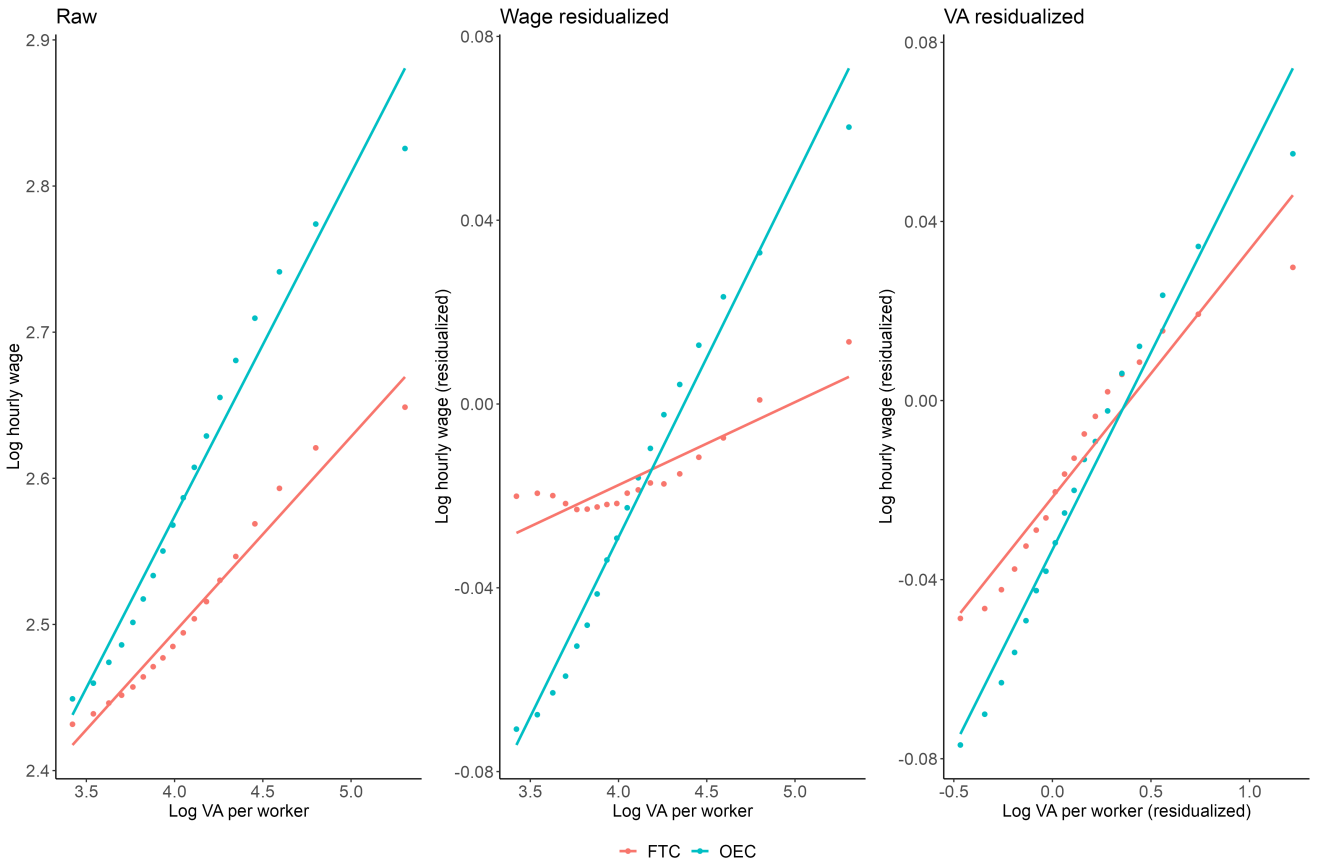


Figure 3.4 – Average firm-level log hourly wage in FTCs and OECs as a function of value added per worker, raw, wage residualized and VA residualized.

Notes: The first panel shows the binscatter of average log hourly wage at the firm-year level against log value added per worker. In the second panel, wages are residualized with respect to year, gender, age, occupation, experience in occupation and industry. In the third panel, log value added per worker is additionally residualized with respect to the same variables. The bottom 10% in terms of value added per worker are trimmed.

Correlation with average FTC length Having explored the predictions of the model, I now turn to another important dimension of FTCs, their length. The model does not give predictions about FTC length, but given that one of the main mechanisms is job security, FTC length is a highly relevant dimension. Indeed, as was illustrated above, there is significant variation in average FTC length, which suggests different practices and uses. As before, Figure 3.5 shows average firm-level log hourly wage in FTCs and OECs as a function of average FTC length in days, both raw and residualized.

Strikingly, once residualized, wages in FTCs decrease with average FTC length, which is in line with the interpretation of compensating differentials for job security. In the last panel, for

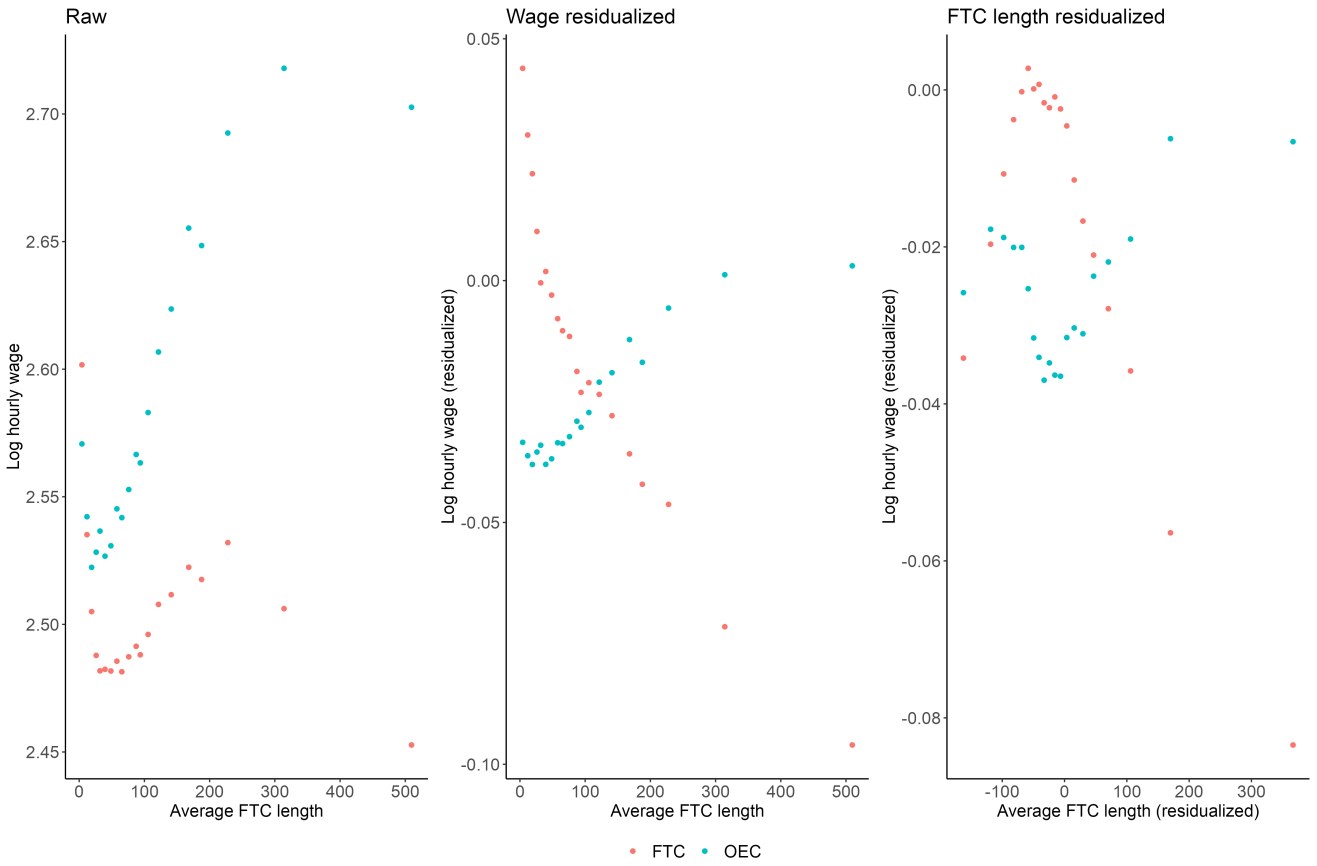


Figure 3.5 – Average firm-level log hourly wage in FTCs and OECs as a function of average FTC length in days, raw and residualized.

Notes: The first panel shows the binscatter of average log hourly wage at the firm-year level against with average FTC length in days. In the second panel, wages are residualized with respect to year, gender, age, occupation, experience in occupation and industry. In the third panel, average FTC length is additionally residualized with respect to the same variables.

average residualized FTC length above zero, the same pattern holds. On the other hand, the slopes reverse for lower values residualized FTC length. The interpretation of this is unclear, but it may be related to the same factors that affect binscatters for low productivity firms mentioned above.

Discussion Importantly, in the appendix Figure 3.11, I show that the FTC-OEC wage gap is still decreasing with residualized value added per worker, within FTC length quartiles. Interestingly, the sign switch of the gap is only present for the third FTC length quartile (and maybe the second), whereas the first and fourth are respectively always positive and always negative. This is consistent with the model for different values of the job security parameter δ. For firms with

very short FTCs, the wage compensation always dominates, whereas the opposite holds true for firms which offer long FTCs.

Another important dimension is industry. The appendix Figure 3.12 shows that the FTC-OEC gap is decreasing in every industry. There are two industries for which the sign of the gap does not switch. In financial and insurance activities, the gap is always negative, whereas for education, human health and social activities[17], the gap is always positive. This makes intuitive sense as for instance, the human health and social activities in France are known to use a lot FTCs, and in particular short-term FTCs. The proposed model therefore also seems to hold within industry. On the other hand, it is also clear that there is a lot of heterogeneity across industries, which could reflect different mechanisms. Finally, in the appendix Figure 3.13 I show the residualized FTC-OEC wage gap by year. The graph is very similar across years, which suggests that the results so far are robust to business cycle effects.

6.2 How are pay premia shared between OEC and FTC workers?

In this section, I present a different way of evaluating the differential rent-sharing hypothesis. I will compare and correlate firm pay premia for FTCs and OECs within firms using the AKM approach pioneered in Card et al. (2016) and used in the context of outsourcing in Drenik et al. (2022). This provides a measure of the degree to which high-wage firms for FTCs are also high-wage firms for OECs. In the main body of the paper, I will implement the approach as described in Kline et al. (2020), but in the appendix I also present the alternative approach implemented in Drenik et al. (2022). The advantages of this approach are that firm pay premia capture aspects of wage setting directly. Additionally, the methodology is better suited to account for sorting patterns of workers to firms.

Methodology As in the previous section, I restrict the analysis to hiring wage, defined as the wages in the first year working for an establishment (aggregated across the first spell for OECs). I then estimate an AKM specification *separately* for OECs and FTCs. Formally, I estimate the following specification:

$$\ln hw_{it} = \alpha_i^C + \psi_{J(it)}^C + X_{it}\beta^C + \varepsilon_{it} \tag{3.3}$$

[17] I excluded public administration and defense from the data

where C is either FTC or OEC. I will discuss an alternative approach with a joint estimation approach below. The α_i^C are individual fixed effects, and $\psi_{J(it)}^C$ are firm-contract type specific fixed effects, where $J(it)$ denotes the firm of worker i at time t. The controls include 2-digit occupation, experience in occupation and age fixed-effects.

The identification hypothesis for this approach are the standard identification assumptions in the AKM, namely strong exogeneity of the residuals. In the context of AKM, this translates to the hypothesis of exogenous mobility of workers between firms, which rules out for instance selection of workers into firms based on the match-specific component (see for instance Card et al. (2016) for a more detailled discussion). It additionally imposes that there be no serial correlation and a strong hypothesis on linearity and symmetry of the effects. This is generally considered a strong assumption, but it is still standard in the literature. In my context, I need to make this assumption separately for OECs and FTCs. It is not a priori evident whether how this affects the credibility of the assumptions.

These regressions are estimated on leave-one out connected sets, which are connected sets of firms which stay connected even when dropping one worker. The restriction to a connected set is a well-known fact in the AKM literature to obtain well-defined firm fixed effects. The additional restriction of a leave-one out connected set comes from Kline et al. (2020) who give a methodology for correcting the variance components of the estimated fixed-effects, even in the presence of heteroskedasticity (henceforth the KSS approach, compared with the standard Plug in approach). This is necessary because variance components are biased due to limited mobility bias. The correction will be necessary for me when I compare the firm fixed effects. This restriction comes at the price of a fairly strong restriction as illustrated in Table 3.2. In particular, the average and variance of the wages is higher in the restricted set.

Table 3.2 – Effect of restriction to the leave-one-out connected set

	FTC		OEC	
	Full	Leave-one-out connected set	Full	Leave-one-out connected set
Number of workers	9,732,262	2,533,388	11,130,752	2,433,793
Number of firms	858,645	336,053	1,026,383	315,309
Number of observations	17,334,367	5,885,385	15,943,203	4,598,481
Average log hourly wage	2.55	2.59	2.68	2.71
Median log hourly wage	2.46	2.48	2.56	2.58
Variance of log hourly wage	0.12	0.15	0.19	0.18

Notes: This table shows summary statistics of the data, before and after the leave-one-out connected set restriction.

Table 3.3 – Variance decomposition for the separate by contract AKM specifications

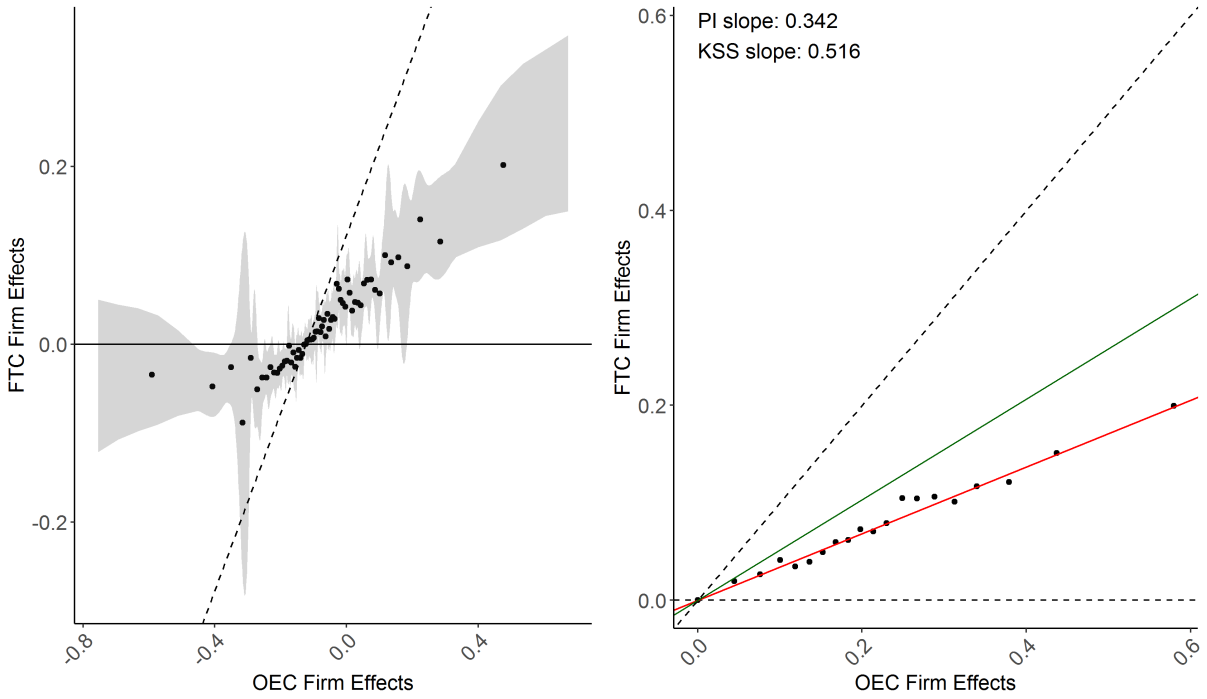
	Fixed-term contracts	Open-ended contracts
Variance of wages	0.15	0.19
Variance of person effects		
Plug in	0.058	0.076
Homoscedasticity only	0.032	0.057
Leave Out	0.037	0.054
Variance of firm effects		
Plug in	0.025	0.019
Homoscedasticity only	0.019	0.014
Leave Out	0.020	0.013
Covariance of firm and person effects		
Plug in	0.005	0.004
Homoscedasticity only	0.009	0.008
Leave Out	0.008	0.009

Notes: This table shows the result of the variance decomposition for wages by contract type. To illustrate the importance of the corrections, it shows the values of the variance decomposition for three different methods. Plug-in corresponds to the uncorrected approach, Homoscedasticity only is the correction by Andrews et al. (2012) and Leave Out is the correction advocated by Kline et al. (2020).

Table 3.3 shows the results from the variance decomposition of the two separate by contract type AKM regressions. The variance is higher for OEC than for FTCs, but interestingly the firm component is significantly more important for FTCs than OECs (13% of the variance versus 7% of the variance). Conversely, person effects are more important for OECs than FTCs.

The estimation above yields a separate firm fixed effect for FTC and OEC workers. Figure 3.6a shows a raw binscatter of the correlation between the two sets of fixed effects. Until an OEC firm effect of -0.35, the relation is constant. In figure 3.6b, I therefore restrict to firm effects above that threshold, bin the observations by OEC firm effect ventiles and then normalize the firm effects so that the lowest ventile has zero firm effects. This does not affect the slope which is the quantity of interest.

As was already noted in Card et al. (2016), the simple regression of FTC firm effects on OEC firm effects is biased because of the errors in the estimation of the firm effects. Here, I follow the approach in Kline et al. (2020) to correct for this error. It simply notices that the bias of



(a) Raw binscatter of OEC firm effects against FTC firm effects.

(b) Restricted plot of OEC firm effects against FTC firm effects, binned by ventiles.

Figure 3.6 – Firm-level pay premia sharing between FTC and OEC workers.

Notes: Panel 3.6a shows binscatter of OEC firm effects against FTC firm effects. Observations are weighted by total number of person-year observations in a firm. The firm effects are estimated separately using a standard AKM specification. The binscatter is produced following the recommendations by Cattaneo et al. (2022) and using their package for inference. The bands are 95% confidence bands. The dashed line is simply for reference of a slope of 1. In panel 3.6b, I restrict to firm after the kink, namely firms with OEC firm effects larger than -0.35. I then bin the observations by OEC firm effect ventiles (weighted by total number of person-year observations) and normalize the lowest ventile to have zero firm effects. The red line represents the estimated OLS slope of the data. The green line represents the slope when correcting for measurement error induced bias by multiplying a correction factor $\frac{\text{var}_{\text{Plug In}}\left(\psi^{\text{OEC}}\right)}{\text{var}_{\text{Leave Out}}\left(\psi^{\text{OEC}}\right)}$.

the regression comes from the denominator of the regression coefficient which involves the variance of the OEC firm effects[18], which is biased. They therefore simply multiply the estimated regression coefficient by a correction factor given by $\frac{\text{var}_{\text{Plug In}}\left(\psi^{\text{OEC}}\right)}{\text{var}_{\text{Leave Out}}\left(\psi^{\text{OEC}}\right)}$. This correction factor comes out to 1.5. The plug in slope is 0.34 and the KSS corrected slope is therefore 0.51. The regressions are reported in the appendix table 3.15.

Discussion Figure 3.6b reports that the OEC and FTC pay premia trace out a slope of 0.51. In other words, if we compare two firms, A and B, and B offers 10% pay premium for its

[18]Indeed, the coefficient in the regression of FTC firm effects on OEC firm effects is $\frac{\text{cov}\left(\psi^{\text{FTC}},\psi^{\text{OEC}}\right)}{\text{var}\left(\psi^{\text{OEC}}\right)}$

OEC workers compared to firm A, then the corresponding pay premium for FTC workers at B compared to A is predicted to be 5.1%. Therefore, firms do seem to share some of their pay premia with FTC workers but only about half what OEC workers get.

There are two benchmarks, which are also represented in Figure 3.6b by the dashed lines. If for instance the market for FTCs was perfectly competitive, then firms would not pay any wage premia to FTC workers, which would yield a slope of zero. On the other hand, if there was no differential rent-sharing at all, then the firm premia would be identical yielding a slope of 1. The result reported above falls between the two.

Most closely related is Drenik et al. (2022) who compare the pay premia of temporary outsourced workers to regular workers. They find a slope of 0.49, very close to the result in this paper. On the other hand, Card et al. (2016) find a slope of 0.89 when comparing pay premia of men and women and Kline et al. (2020) find a slope of 0.98 when comparing the pay premia of younger and older workers.

Robustness As a robustness check, I also implement the alternate methodology in Drenik et al. (2022). They estimate a modified AKM specification, which allows for the firm fixed effect to vary by contract type. Formally, I estimate the following specification:

$$\ln hw_{it} = \alpha_i + \psi_{J(it)}^{\mathrm{C}_{it}} + \varepsilon_{it} \tag{3.4}$$

where α_i are worker fixed effects, and $\psi_{J(it)}^{\mathrm{CDD}_{it}}$ are firm-contract type specific fixed effects. The superscript C_{it} indicates the contract type the worker i is employed in at time t, and $J(it)$ denotes the firm of worker i at time t. For purely computational reasons, I do not include controls in this specification[19].

It is intuitively appealing to estimate the regression jointly. One important difference to the methodology above is that the individual fixed effect is not differentiated by contract type anymore. This a priori a more constraining hypothesis than the previous approach (especially given the discussion in the literature about productivity differences between contracts). On the other hand, simultaneous estimation should have more power. Ultimately, I chose this method for robustness because the econometric properties of this approach are less clear. In particular, the bias-correction applied above is not readily adaptable. Instead, Drenik et al. (2022) apply a

[19] In robustness not reported here, I checked that the omission of controls only had a very limited effect

split-sample IV approach to correct for the measurement error bias. It consists in splitting the data in half randomly and estimating the fixed effects separately, and then instrumenting the OEC firm effects in one split by the OEC firm effects in the other split. While intuitive and often used (see for instance Babet et al. (2022) or Bassier et al. (2022)), there are no well-established econometric guarantees.

I report the result from these regression in table 3.16, which shows an uncorrected slope of 0.36, very close to the previous estimate of 0.34. The corrected slope, corrected using the split-sample IV approach, is 0.42, somewhat lower than the result in the main section, but the result of significant differential rent-sharing stands.

7 Local labor market heterogeneity

In this section, I explore whether variation in concentration in local labor market generates variation in differential rent-sharing as is predicted by prediction 4. This is predicated on the idea that concentration can be used as a proxy for labor market power. This approach is now standard in the literature (see for example Marinescu et al. (2021), Azar et al. (2020)), but it has also been repeatedly emphasized that concentration is an imperfect proxy for labor market power (for instance Bassier et al. (2022) find small correlations between concentration and labor market power).

I first describe how I measure local labor market concentration, using Labor Market Herfindahl-Hirschman Index (HHI). A novelty in this paper is that I consider concentration separately for FTCs and OECs following the hypothesis that these two markets are relatively segmented. The results in this section suggest that this is a reasonable hypothesis. I then compute the slope of wage-value-added curve by FTC and OEC HHI quartile.

7.1 Measuring labor market concentration

I define the local labor market level as the interaction between a commuting zone and an occupation. I measure hiring concentration by a Labor Market Herfindahl-Hirschman Index (HHI) in hiring as a proxy for labor market power. I follow the standard definition of the HHI (see for instance Azar et al. (2020) or Marinescu et al. (2021)) with two modifications. First, I compute an aggregated HHI over the 2010 to 2014 period (rather than yearly). I do this to

smooth over some of the idiosyncratic time variation. Second, since part of the argument of the model is that FTC and OEC labor markets are segmented, I also compute the HHIs separately by contract type.

I denote the set of firms hiring in occupation o in commuting zone z at time t by $\mathcal{J}_{o,z,t}$. The number of workers hired by firm j is denoted $\mathcal{N}_{j,o,z,t}$. The firm's (aggregated) labor market $s^L_{j,o,m}$ share is then given by:

$$s^L_{j,o,z} = \frac{\sum_t \mathcal{N}_{j,o,z,t}}{\sum_t \sum_{k \in \mathcal{J}_{o,z,t}} \mathcal{N}_{k,o,z,t}}$$

The labor market HHI, $\text{HHI}_{o,z}$, is then given by the sum of these shares squared over all firms in the LLM:

$$\text{HHI}_{o,z,t} = \sum_t \sum_{j \in \mathcal{J}_{o,z,t}} s^2_{j,o,z}$$

Table 3.4 shows summary statistics for the distribution of the HHIs by contract type. Rows 3 and 4 of the table show the same summary statistics after the merge with the firm-level data. I discuss this further below. The median HHI for FTCs is at 0.07 which is similar to what Marinescu et al. (2021) find (for global HHI). [20] Interestingly, the median HHIs for OECs is a bit lower. This holds for both median and mean, before and after merge. The fact FTC markets are more concentrated than OECs is another mild confirmation for our hypothesis that $\eta^{\text{OEC}} > \eta^{\text{FTC}}$.

Table 3.4 – Summary of the FTC and OEC labor market HHI distribution.

	Minimum	Q1	Median	Mean	Q3	Maximum
HHI FTC	0.001	0.034	0.076	0.190	0.198	1.000
HHI OEC	0.000	0.019	0.048	0.159	0.136	1.000
HHI FTC after merge	0.001	0.006	0.012	0.028	0.033	1.000
HHI OEC after merge	0.000	0.003	0.007	0.018	0.019	1.000

Notes: This table shows summary stastics for the distributions of aggregated labor market hiring HHI over the period 2010 to 2014 at local labor market level defined as commuting zone and occupation. Rows 3 and 4 show the same summary statistics after the merge with firm-level data, which is necessary for the results below.

The table above already shows a difference between the distributions of the HHI by contract type. Additionally, the correlation between the two is 0.51 (appendix Figure 3.14 also shows this correlation), which while significant, still leaves a lot of room for separate variation, which is

[20] They use 4-digit occupation codes.

Table 3.5 – Cross-table of firm-occupation observation by FTC and OEC HHI quartiles.

		HHI OEC			
		Q1	Q2	Q3	Q4
HHI FTC	Q1	99,724	38,685	8,677	1,028
	Q2	40,690	60,457	36,905	9,779
	Q3	3,762	35,705	70,078	38,392
	Q4	4,643	12,261	32,431	98,588

Notes: This table shows the number of firm cross occupation observations by FTC and OEC HHI.

consistent at least partially with the hypothesis that the two markets are segmented. Table 3.5 additionally confirms this by showing the number of firm cross occupation observations across FTC and OEC quartiles: the correlation is clear, but there is also large variation when fixing one HHI. All of these observations are consistent with an at least partial segmentation of these two markets.

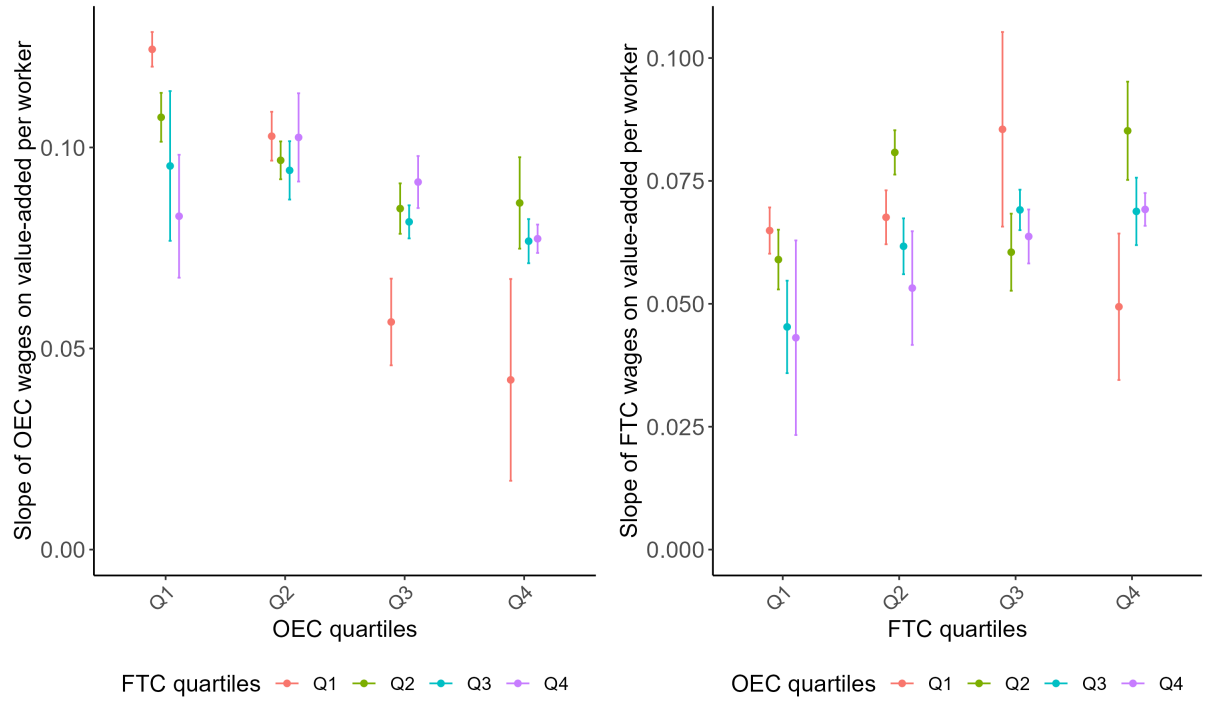
7.2 Results

The goal of this section is to verify prediction 4, which states that decreases in η^C should decrease the corresponding slope between average wages and value-added per worker. I therefore check whether increases in concentration, which correspond to decreases in η^C, satisfy this prediction. Because of the correlation between the FTC and OEC HHI measures I will show the variation in the slope when varying the corresponding HHI quartile and holding the other HHI quartile constant.

To match with the local labor market definition, I collapse the data to the firm and occupation level. Namely, I compute average wages in FTCs and OECs at the firm and occupation level. As before, I residualize wages on year, gender, age, experience in occupation and industry. I then merge the HHI information. I also merge the financial data, and in particular value-added per worker information. Notice that this measure of value-added per worker is constant at the firm-level and will therefore be identical across the occupation cells within a firm. As before, I also residualize value-added per worker.

Figure 3.7 shows the slopes of log hourly wages (FTC wages in panel 3.7b and OEC wages in

panel 3.7a) against log value-added per worker, by FTC and OEC HHI quartiles. As before, both wages and value-added are residualized.



(a) Slopes of log hourly OEC wage against log value-added per worker, both residualized, by OEC and FTC HHI quartiles.

(b) Slopes of log hourly FTC wage against log value-added per worker, both residualized, by OEC and FTC HHI quartiles.

Figure 3.7 – Slopes of log hourly wages against log value-added per worker, both residualized, by OEC and FTC HHI quartiles.

Notes: The figure on the left shows the slope coefficient from a regression of residualized value-added per worker on residualized log hourly wages of OECs, by OEC HHI quartiles on the horizontal axis and FTC HHI quartiles in colors. The residualization variables are year, gender, age, experience in occupation and industry. The regression is computed at the firm cross occupation level. Standard errors are clustered at the firm level. The right graph replaces OEC wages by FTC wages, and the variation along the horizontal axis is now FTC HHI quartiles and the colors are OEC HHI quartiles.

For OECs, prediction 4 seems to be confirmed. Across all FTC quartiles, the slope decreases when the OEC quartile increases. On the other hand, for FTCs, the slopes seem to be fairly constant, if not even slightly increasing.

7.3 Discussion

The fact that prediction 4 bears out for OECs but FTCs is an interesting observation. I give a few tentative explanations below, but further exploration is left for future research.

As was discussed above, the labor supply elasticities, and therefore ultimately the labor market power, can come from many different sources. In particular, it is possible that the dominant sources vary between the OEC and FTC markets. The results above suggest that concentration does play a significant role in labor market power for OECs, whereas it does not for FTCs. This is consistent for instance with the hypothesis that for FTCs, other sources are more important such as lower bargaining power (maybe due to less union coverage). It is also possible that variation in concentration capture different aspects of labor markets for different contract types (for instance outside options, see Bassier et al. (2022) for a related discussion).

8 Conclusion

In this paper, I explore the effect of fixed-term contracts on the wage structure, both theoretically and empirically. I start by providing a monopsonistic model of segmented labor markets categorized by contract type, augmented by the notion of compensating wage differentials for job security.

I start by documenting that the average wage gap is precisely estimated to be zero. This strongly suggests the existence of mechanisms which have opposing effects on the sign of the wage gap, such as compensating wage differentials and differential rent-sharing in my model.

I then document significant differential rent-sharing between contract types, using both a descriptive approach and an approach based on the estimation of firm wage premia. This is in line with the mechanisms advocated in the model. I additionally document that rent-sharing varies with local labor market concentration for OECs, but not for FTCs.

The results in this paper open interesting avenues for future research. From a theoretical perspective, a microfounded general equilibrium model would allow to more cleanly identify consequences for policy. Empirically, it would be important to further understand the various rationales and mechanisms underlying the heterogeneity in the use of FTCs and how they affect wages. In particular, the role of management and firm organization is an interesting avenue.

A Appendix

A.1 Additional figures

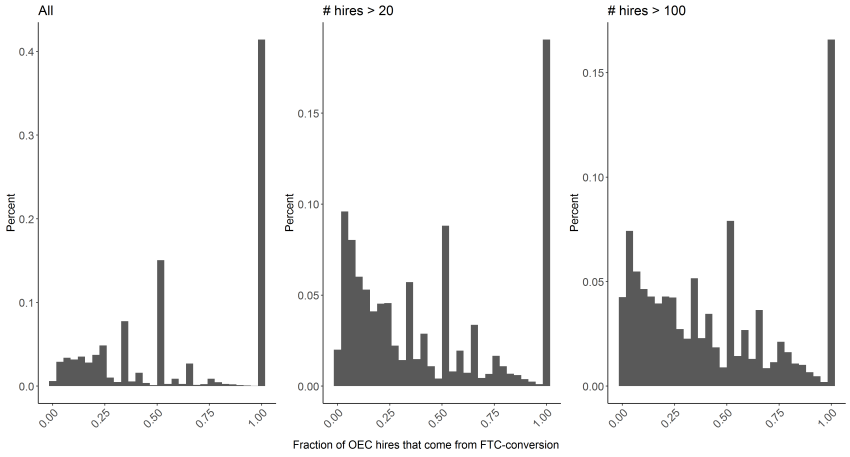


Figure 3.8 – Distribution of the fraction of OECs hired in a year that were previously employed in an FTC at the firm

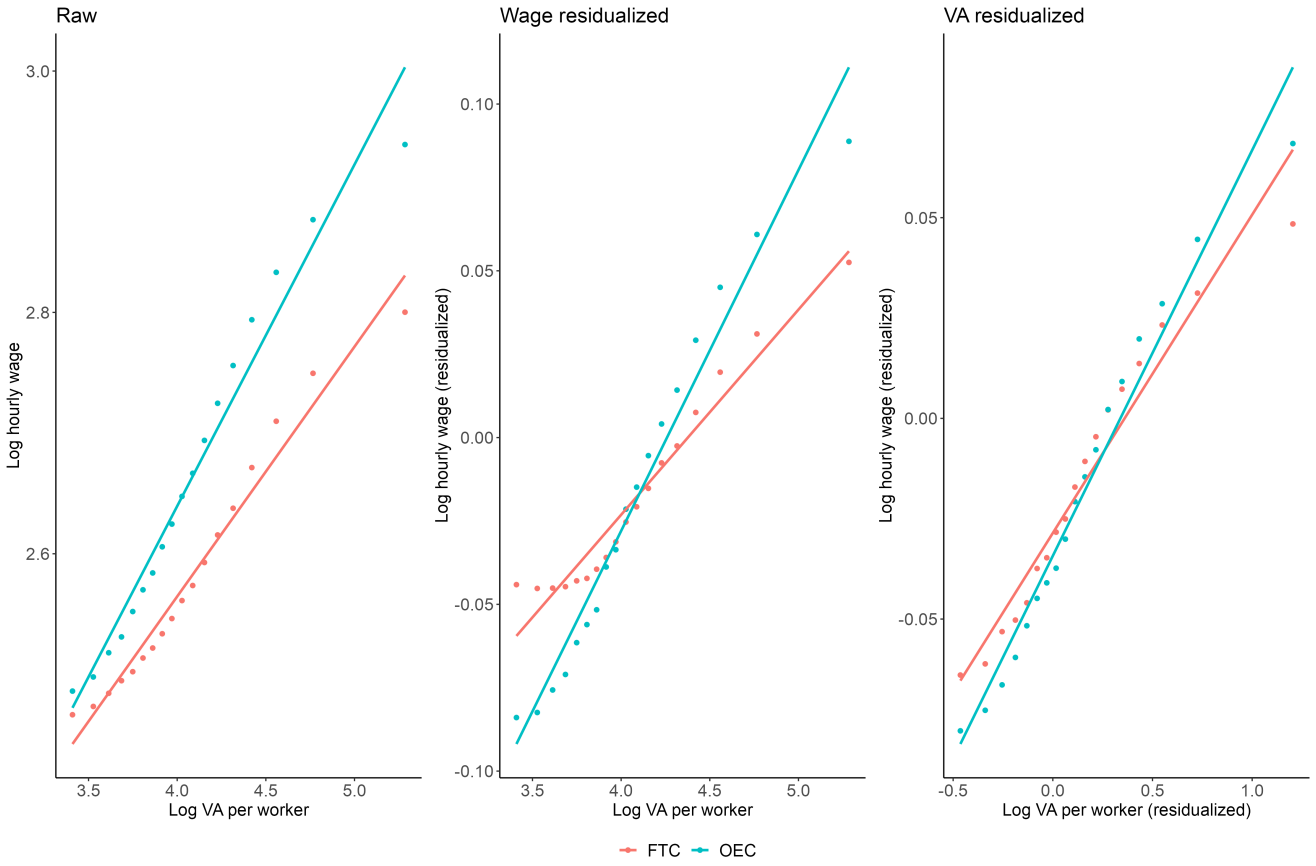


Figure 3.9 – Average firm-level log hourly wage in FTCs and OECs as a function of value added per worker, raw, wage residualized and VA residualized and additionally restricted to workers aged over 30.

Notes: The first panel shows the binscatter of average log hourly wage at the firm-year level against with log value added per worker. In the second panel, wage are residualized with respect to year, gender, age, occupation, experience in occupation and industry. In the third panel, log value added per worker is additionally residualized with respect to the same variables. The bottom 10% in terms of value added per worker are trimmed. The sample is restricted to individuals aged over 30.

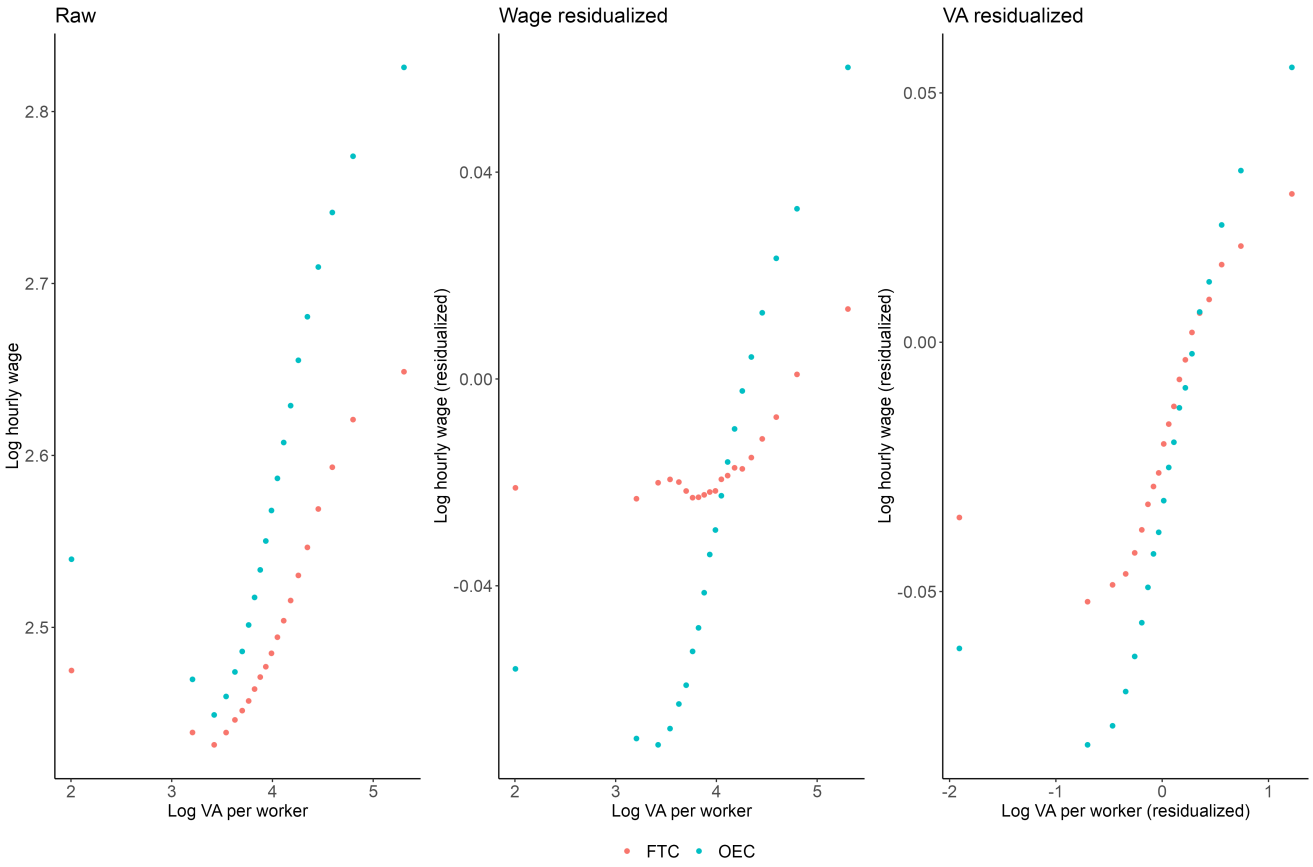


Figure 3.10 – Average firm-level log hourly wage in FTCs and OECs as a function of value added per worker, raw, wage residualized and VA residualized. Full sample

Notes: The first panel shows the binscatter of average log hourly wage at the firm-year level against with log value added per worker. In the second panel, wage are residualized with respect to year, gender, age, occupation, experience in occupation and industry. In the third panel, log value added per worker is additionally residualized with respect to the same variables.

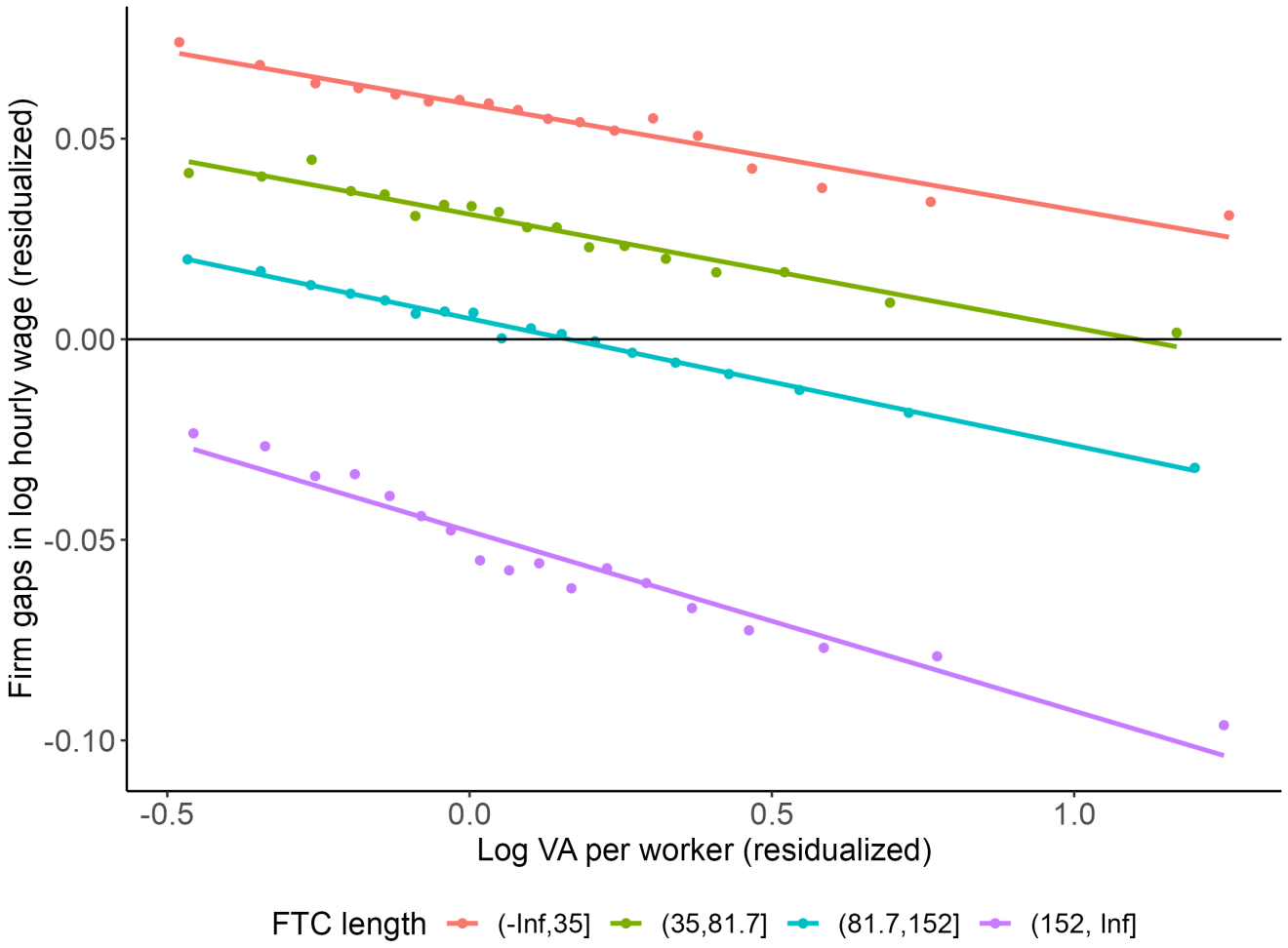


Figure 3.11 – Average residualized firm-level log hourly wage in FTCs and OECs as a function of residualized value added per worker, by average FTC length quartiles.

Notes: The figures shows residualized average firm-level FTC-OEC wage gap as a function of residualized value added per worker, by average FTC length quartiles.

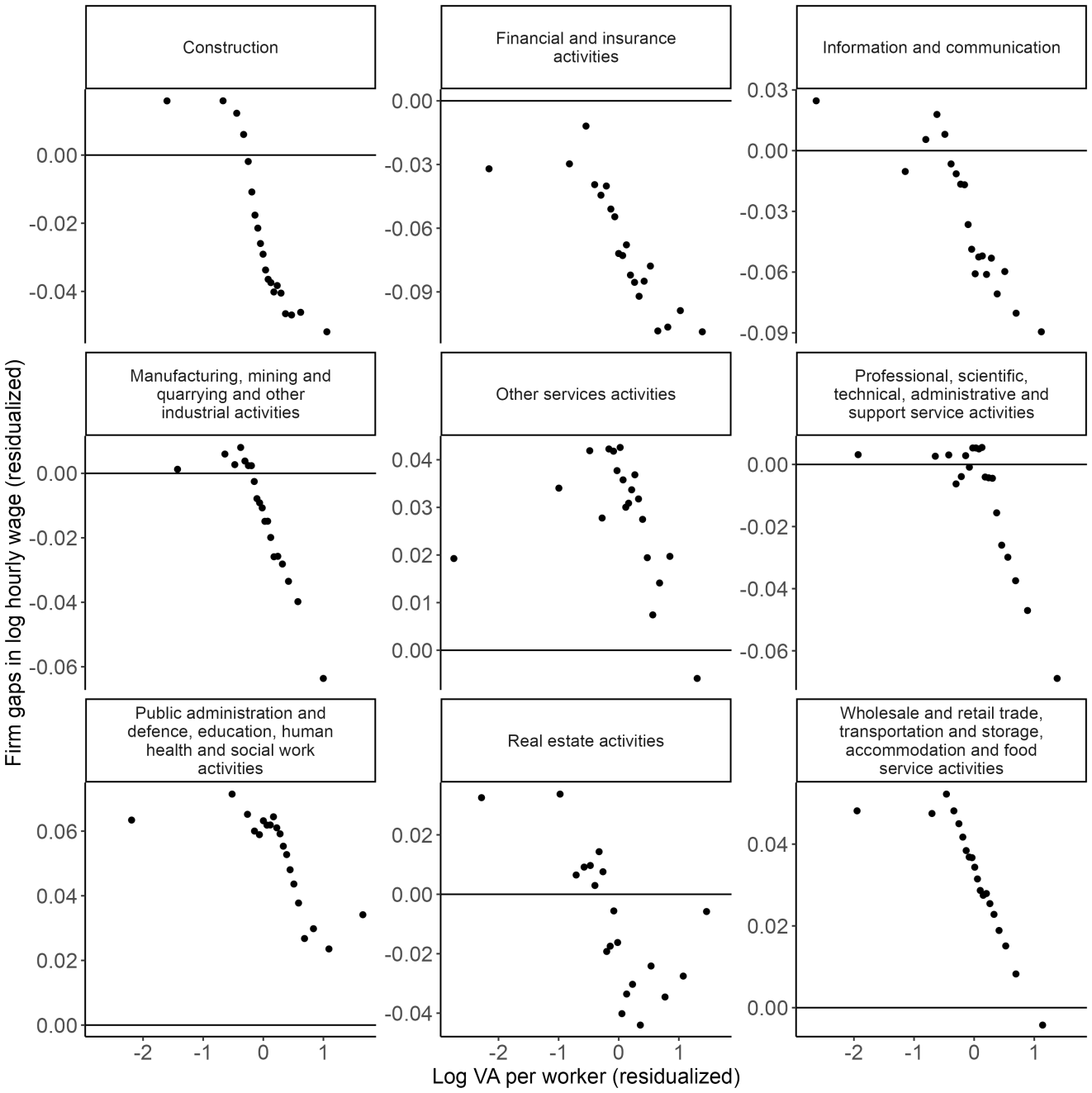


Figure 3.12 – Residualized average firm-level FTC-OEC wage gap as a function of residualized value added per worker, by industry.

Notes: This shows residualized average firm-level FTC-OEC wage gap as a function of residualized value added per worker, by industry. Industry is defined at a broad level with 10 different industries.

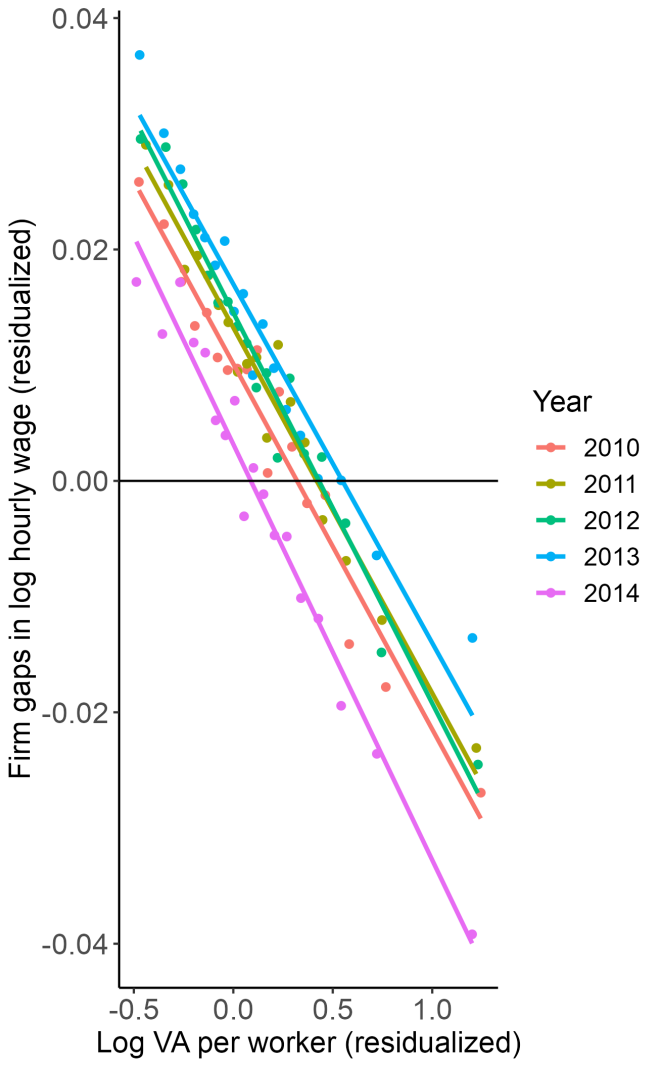


Figure 3.13 – Residualized average firm-level FTC-OEC wage gap as a function of residualized value added per worker, by year.

Notes: This shows residualized average firm-level FTC-OEC wage gap as a function of residualized value added per worker, by year.

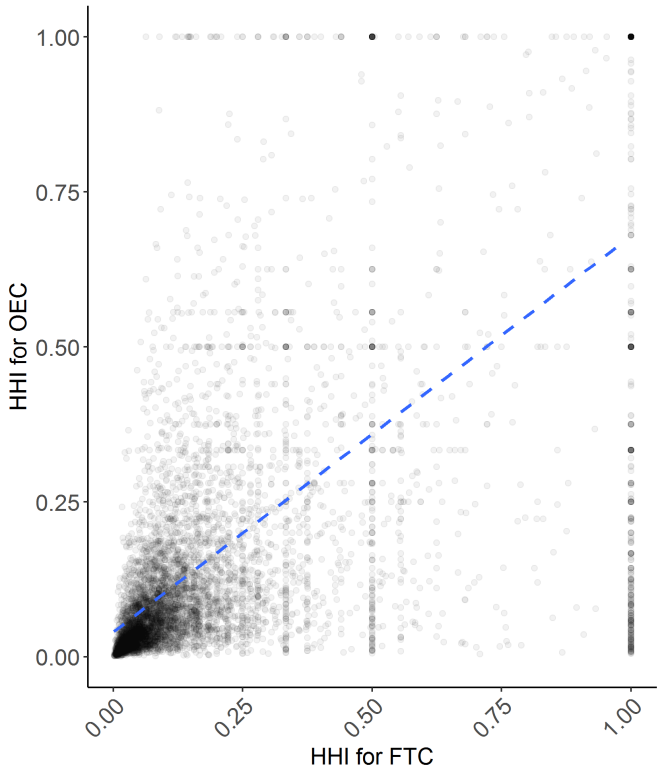


Figure 3.14 – Plot of FTC HHIs against OEC HHIs.

Notes: This figure is a scatter plot of local labor market HHIs for FTCs and OECs. The blue line is the slope from an OLS regression.

A.2 Additional tables

	English description	French description (short)	French description (long)
10	Farmers	Agriculteurs	Exploitants de l'agriculture, sylviculture, pêche et aquaculture
21	Craftsmen	Artisans	Artisans
22	Shopkeepers and similar	Commerçants et assimilés	Commerçants et assimilés
23	Entrepreneurs with more than 10 employees	Chefs d'entreprise de plus de 10 personnes	Chefs d'entreprise de plus de 10 personnes
31	Liberal professions	Professions libérales	Professions libérales
33	Civil service executives	Cadres de la fonction publique	Cadres administratifs et techniques de la fonction publique
34	Teachers and scientists	Professeurs et professions scientifiques	Professeurs et professions scientifiques supérieures
35	Information, arts and entertainment professionals	Professions de l'information, de l'art et des spectacles	Professions de l'information, de l'art et des spectacles
37	Administrative and sales executives	Cadres administratifs et commerciaux	Cadres des services administratifs et commerciaux d'entreprise
38	Technical business executives	Cadres techniques d'entreprise	Ingénieurs et cadres techniques d'entreprise
42	Primary and vocational education professions	Professions de l'enseignement primaire et professionnel	Professions de l'enseignement primaire et professionnel, de la formation continue et du sport
43	Health and social work professionals	Intermédiaires de la santé et du travail social	Professions intermédiaires de la santé et du travail social
44	Religious	Religieux	Ministres du culte et religieux consacrés
45	Intermediate civil servants	Intermédiaires de la fonction publique	Professions intermédiaires de la fonction publique (administration, sécurité)
46	Business intermediaries	Intermédiaires des entreprises	Professions intermédiaires administratives et commerciales des entreprises
47	Technicians	Techniciens	Techniciens
48	Production supervisors	Agents de maîtrise de production	Agents de maîtrise (hors maîtrise administrative)
52	Civil servants	Employés de la fonction publique	Employés administratifs de la fonction publique, agents de service et auxiliaires de santé
53	Police, military and private security officers	Policiers, militaires et agents de sécurité privée	Policiers, militaires, pompiers, agents de sécurité privée
54	Company administrative employees	Employés administratifs d'entreprise	Employés administratifs d'entreprise
55	Commercial employees	Employés de commerce	Employés de commerce
56	Personal service workers	Personnels des services aux particuliers	Personnels des services directs aux particuliers
62	Skilled industrial workers	Ouvriers qualifiés de type industriel	Ouvriers qualifiés de type industriel
63	Skilled craft workers	Ouvriers qualifiés de type artisanal	Ouvriers qualifiés de type artisanal
64	Transport drivers	Conducteurs du transport	Conducteurs de véhicules de transport, chauffeurs-livreurs, coursiers
65	Machine and warehouse operators	Conducteurs d'engins et magasiniers	Conducteurs d'engins, caristes, magasiniers et ouvriers du transport (non routier)
67	Low-skilled industrial workers	Ouvriers peu qualifiés de type industriel	Ouvriers peu qualifiés de type industriel
68	Low-skilled craft workers	Ouvriers peu qualifiés de type artisanal	Ouvriers peu qualifiés de type artisanal
69	Agricultural workers	Ouvriers agricoles	Ouvriers agricoles, des travaux forestiers, de la pêche et de l'aquaculture

Table 3.6 – Description of 2-digit occupation codes

Table 3.7 – FTC-OEC hiring wage gap regression, controlling for 2-digit occupation codes rather than 4-digit.

	Log hourly wage							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
FTC	-0.1311***	-0.1215***	-0.0673***	-0.0701***	-0.0006	0.0096***	-0.0045***	0.0265***
	(0.0022)	(0.0022)	(0.0021)	(0.0014)	(0.0010)	(0.0009)	(0.0006)	(0.0005)
Observations	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570	33,277,570
R^2	0.02689	0.04358	0.12355	0.27117	0.50588	0.51625	0.66253	0.87947
Within R^2	0.02682	0.02333	0.00747	0.00879	9.57×10^{-7}	0.00023	4.29×10^{-5}	0.00141
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓	✓	✓
Industry fixed effects				✓	✓	✓	✓	✓
Occupation (2 digit) fixed effects					✓	✓	✓	✓
Exp. in occ. fixed effects						✓	✓	✓
Establishment fixed effects							✓	✓
Individual fixed effects								✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%, and FTC wages are aggregated over the FTC spell. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 2-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.8 – FTC-OEC hiring wage gap regression for the 2005 to 2009 period

	Log hourly wage							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
FTC	-0.1410***	-0.1277***	-0.0615***	-0.0702***	-0.0007	0.0043***	-0.0042***	0.0243***
	(0.0022)	(0.0021)	(0.0021)	(0.0015)	(0.0010)	(0.0010)	(0.0006)	(0.0007)
Observations	37,213,647	37,213,647	37,213,647	37,213,647	37,213,647	37,213,647	37,213,647	37,213,647
R^2	0.02703	0.04802	0.13030	0.26505	0.54538	0.55049	0.67913	0.89014
Within R^2	0.02698	0.02249	0.00540	0.00752	1.02×10^{-6}	4.44×10^{-5}	3.62×10^{-5}	0.00113
Year fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓	✓	✓
Industry fixed effects				✓	✓	✓	✓	✓
Occupation (2 digit) fixed effects					✓	✓	✓	✓
Exp. in occ. fixed effects						✓	✓	✓
Establishment fixed effects							✓	✓
Individual fixed effects								✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2005 to 2010. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%, and FTC wages are aggregated over the FTC spell. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.9 – FTC-OEC hiring wage gap regression over the period 2010 to 2014 using cross-sectional DADS postes data

	Log hourly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
FTC	-0.1330***	-0.1229***	-0.0680***	-0.0776***	-0.0132***	-0.0186***
	(0.0024)	(0.0024)	(0.0024)	(0.0015)	(0.0007)	(0.0006)
Observations	33,745,136	33,745,136	33,745,136	33,745,136	33,745,136	33,745,136
R^2	0.02553	0.04265	0.12141	0.26967	0.52699	0.65580
Within R^2	0.02541	0.02198	0.00702	0.00987	0.00039	0.00067
Year fixed effects	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓
Industry fixed effects				✓	✓	✓
Occupation fixed effects					✓	✓
Establishment fixed effects						✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%. Compared to table 3.1 the data used is cross-sectional DADS postes data, rather than the statistical match panel data. This also means that FTC wages are *not* aggregated over the FTC spell (see discussion in section 4.3). Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.10 – FTC-OEC hiring wage gap regression over the period 2010 to 2014 with FTC wages not aggregated over the spell.

	Log hourly wage					
	(1)	(2)	(3)	(4)	(5)	(6)
FTC	-0.1387***	-0.1291***	-0.0750***	-0.0792***	-0.0136***	-0.0186***
	(0.0022)	(0.0022)	(0.0021)	(0.0014)	(0.0007)	(0.0006)
Observations	33,277,572	33,277,572	33,277,572	33,277,572	33,277,572	33,277,572
R^2	0.02987	0.04654	0.12600	0.27194	0.54437	0.67216
Within R^2	0.02978	0.02612	0.00921	0.01112	0.00047	0.00076
Year fixed effects	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓
Industry fixed effects				✓	✓	✓
Occupation fixed effects					✓	✓
Establishment fixed effects						✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%. Compared to table 3.1, FTC wages are *not* aggregated over the FTC spell (see discussion in section 4.3). This is done to facilitate the comparison with table 3.9 which uses cross-sectional data and therefore cannot do that adjustment. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.11 – FTC-OEC hiring wage gap regression over the period 2010 to 2014, restricted to individuals aged 30 to 60.

	Log hourly wage						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
FTC	-0.1239***	-0.1138***	-0.1113***	-0.1079***	-0.0141***	-0.0027***	0.0067***
	(0.0031)	(0.0030)	(0.0030)	(0.0020)	(0.0008)	(0.0008)	(0.0006)
Observations	17,407,406	17,407,406	17,407,406	17,407,406	17,407,406	17,407,406	17,407,406
R^2	0.01889	0.04379	0.04906	0.26643	0.58327	0.58937	0.72328
Within R^2	0.01868	0.01613	0.01551	0.01598	0.00043	1.56×10^{-5}	8.41×10^{-5}
Year fixed effects	✓	✓	✓	✓	✓	✓	✓
Sex fixed effects		✓	✓	✓	✓	✓	✓
Age fixed effects			✓	✓	✓	✓	✓
Industry fixed effects				✓	✓	✓	✓
Occupation fixed effects					✓	✓	✓
Exp. in occ. fixed effects						✓	✓
Establishment fixed effects							✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014. Individuals are restricted to ages between 30 and 60 years old. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%, and FTC wages are aggregated over the FTC spell. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.12 – FTC-OEC hiring wage gap regression over the period 2010 to 2014, by industries which can potentially use CDD-U.

	Log hourly wage		
CDD-U industries	Full sample	FALSE	TRUE
	(1)	(2)	(3)
FTC	0.0027***	-0.0034***	0.0053***
	(0.0007)	(0.0006)	(0.0009)
Observations	22,241,559	5,746,543	16,495,016
R^2	0.68068	0.67568	0.67818
Within R^2	1.96×10^{-5}	4.79×10^{-5}	6.76×10^{-5}
Year fixed effects	✓	✓	✓
Sex fixed effects	✓	✓	✓
Age fixed effects	✓	✓	✓
Industry fixed effects	✓	✓	✓
Occupation fixed effects	✓	✓	✓
Exp. in occ. fixed effects	✓	✓	✓
Establishment fixed effects	✓	✓	✓

Notes: This table reports the result from the regression of log hourly wage on an indicator for whether the observation is in an FTC, and a set of fixed-effects over the period 2010 to 2014 over the full sample, and for two sets of industries depending on their right to use CDD-U. The sample is different from 3.1 because it is the sample that is used in the heterogneity analysis, after matching with financial and other firm data. As can be seen from the the estimate on the full sample, the estimates are nearly identical. Log hourly wage is computed as gross wage divided by number of hours, inflation-adjusted into 2012 euros, trimmed at 0.1%, and FTC wages are aggregated over the FTC spell. Age is separated into 5 year bins. Industry is a 38 level code for industries. Occupation is the 4-digit PCS code. Experience in occupation is the number of years since the individual first practiced the 2-digit occupation of the observation. The results are clustered at the establishment level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.13 – Firm selection when creating the firm-level dataset

Restriction	Number of observations	Number of firms	Mean firm size	Mean log hourly wage
Full data	3231096	1241254	10.32	2.57
FTC-OEC gap	1105368	535224	23.31	2.56
Financial data	937934	452965	22.61	2.56
Hiring information	895988	434706	23.16	2.55
Missing FTC length information	815437	400756	24.49	2.56
Missing VA information	784633	384635	24.64	2.55

Notes: This table shows the regression log hourly wage for FTCs and OECs on log value added per worker, restricted to the range of value added per worker above 3.5 identified manually from figure 3.4. Columns (1) and (2) are for raw wages, whereas (3) and (4) are residualized with respect to year, gender, age, occupation, experience in occupation and industry. Standard errors are clustered at the firm level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.14 – Regression of log hourly wage on log value added per worker, restricted to the appropriate value added per worker range.

	Log hourly wage		Residualized Log hourly wage			
	FTC	OEC	FTC	OEC	FTC	OEC
	(1)	(2)	(3)	(4)	(5)	(6)
Log Value Added per worker	0.1326***	0.2319***	0.0188***	0.0774***		
	(0.0009)	(0.0012)	(0.0006)	(0.0007)		
Residualized Log Value Added per worker					0.0545***	0.0855***
					(0.0007)	(0.0007)
Observations	687,912	687,912	687,912	687,912	690,357	690,357
R^2	0.06846	0.12787	0.00230	0.03748	0.01536	0.03620
Adjusted R^2	0.06846	0.12787	0.00230	0.03748	0.01536	0.03620

Notes: This table shows the regression log hourly wage for FTCs and OECs on log value added per worker, restricted to the same range of value added per worker as in figure 3.4. Columns (1) and (2) are for raw wages, whereas (3) and (4) are residualized with respect to year, gender, age, occupation, experience in occupation and industry and (5) and (6) log value added per worker is residualized with respect to the same variables. Standard errors are clustered at the firm level.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.15 – Regression of FTC firm effects on OEC firm effects.

	FTC-FFE	
	(1)	(2)
(Intercept)	0.0440***	0.0445***
	(0.0018)	(0.0018)
OEC-FFE	0.3082***	0.3387***
	(0.0134)	(0.0155)
Observations	148,916	140,911
R^2	0.08380	0.08799
Adjusted R^2	0.08379	0.08799

Notes: This table reports the regression of FTC firm effects on OEC firm effects, weighted by total number of person-year observations by firm. The firm effects are estimated separately using a standard AKM specification. Column (1) reports the regression on the full data. Column (2) restricts to firms with an OEC firm effect greater than -0.35. This threshold is chosen manually to correspond to the linear section in Figure 3.6a.
*p<0.1; **p<0.05; ***p<0.01.

Table 3.16 – Regressions of FTC firm effect on OEC firm effects, estimated jointly and restricted after the kink.

	FTC firm effect		FTC firm effect, 1st split	
	(1)	(2)	(3)	(4)
(Intercept)	0.0100***	0.0117***	-0.1107***	-0.1019***
	(0.0013)	(0.0015)	(0.0031)	(0.0036)
OEC firm effect	0.3702***	0.3803***		
	(0.0124)	(0.0163)		
OEC firm effect, 1st split			0.3512***	0.4190***
			(0.0150)	(0.0194)
Observations	184,695	63,463	63,463	63,463
R^2	0.14929	0.17817	0.14002	0.13479
Adjusted R^2	0.14929	0.17816	0.14000	0.13478

Notes: This table reports the regression of FTC firm effects on OEC firm effects, weighted by total number of person-year observations by firm. The firm effects are estimated jointly without controls. I also split the data in two randomly stratified at the individual level. Column (1) reports the regression on the full data, restricting to firms with an OEC firm effect greater than -0.35. Column (2) estimates the same regression on the subsample of firms which get selected for the leave-out connected set on the first random subsample. Column (3) reports the regression of FTC firm effect in the first split on the OEC firm effect in the first split. Column (4) finally reports the result of the previous regression where I instrument the OEC firm effect in the first split by the OEC firm effect in the second split. This is done to correct for the measurement error bias. In both column (3) and (4), the observations have been restricted to have OEC firm effects (in the first split) larger than -0.2, a threshold identified manually in a binscatter (not reported in the paper).
*p<0.1; **p<0.05; ***p<0.01.

A.3 Wage dynamics in the presence of FTCs

A thorough discussion of wage dynamics is outside the scope of this paper. In lieu of this, I will simply provide the results of a basic event-study approach. The goal is to illustrate the potential existence of the dual wage ladder mentioned above and encourage further investigation into this topic.

Methodology Instead of restricting to hiring wages, I now use the full matched DADS panel as described in section 4.3. For workers with several jobs, I additionally restrict to the dominant job which is I define as the job with the highest hourly wage.[21] This restriction is necessary to define the transitions from FTCs to OECs but it is also restrictive as it for instance drops the wages of non-dominant wages. I then restrict to individuals that have at least one transition from an FTC to an OEC, and for the few instances with more than one such switch I restrict to the first switch.

I then run an event-study *without* a control group:

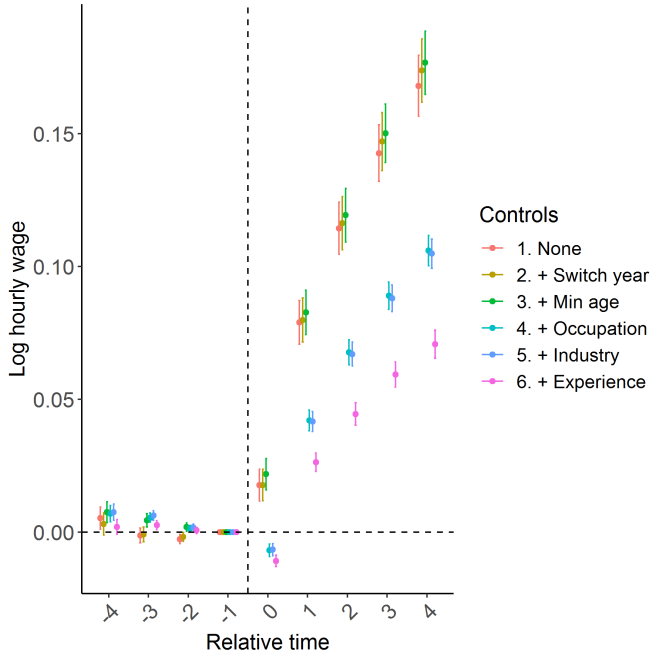
$$\ln hw_{it} = \sum_{k \neq 0} \mathbb{1}[\text{Year}_t - \text{Year\_transition}_i = k] + X_{it} + \varepsilon_{it}$$

where the controls X_{it} include a year-of-switch fixed effect, minimum age over the period of the worker, occupation, industry and second order polynomial of within-firm experience.

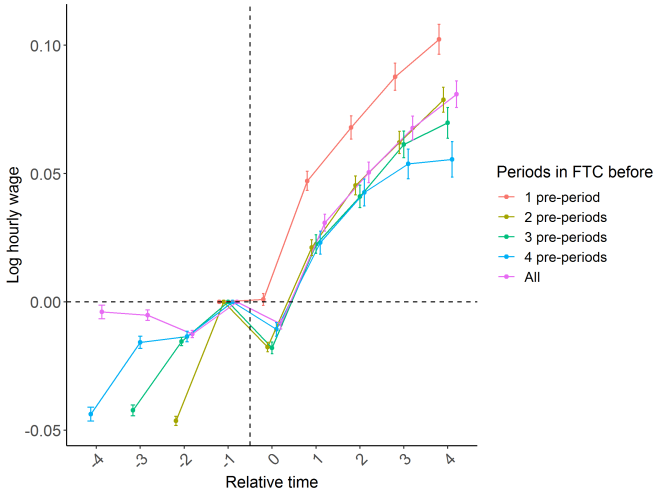
Results In figure 3.15a, I present the event-study around an FTC to OEC without any restrictions. The figure has two interesting features. First, it suggests little to no wage growth while on FTC and significant wage growth once on an OEC. Second, it suggests that on average workers take a wage cut when switching from an OEC to an FTC. Importantly, the wage cut only appears once occupation has been controlled for. Figure 3.15b changes this picture slightly. It restricts to individuals with no career gap (generally unemployment) around the transition and to individuals which have not had an OEC before the transition. These restrictions are meant to deal with possible confounders due to irregular careers and focus only on individuals that have only been on FTCs and then switch to an OEC. The figure additionally mitigates composition effects by disaggregating depending on the number of years in FTC before the transition. The specification used is the one which controls for year-of-switch fixed effect, minimum age over

[21] For the small fraction of individuals for which this doesn't result in a single job, I further use the number of days and then number of hours.

the period of the worker, occupation, industry and second order polynomial of within-firm experience. As opposed to the previous figure, individuals with different numbers of years before the transition experience wage growth on FTCs, as opposed to when they are all grouped together, which is indicative of a composition effect. It remains true that individuals on average experience a wage cut when transitioning from FTCs to OECs.



(a) Event study for individuals transitioning from FTCs to OECs



(b) Event study for individuals transitioning from FTCs to OECs, restricted to individuals without a career interruption and no instance of an OEC before the transition.

Figure 3.15 – Event studies around FTC to OEC transition

Bibliography

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 2003, *113* (2), 231 – 263.

_ , Jiaying Gu, and Shu Shen, "Instrumental Variable Estimation with First-Stage Heterogeneity," *Working paper*, 2022.

Abowd, John M., Francis Kramarz, and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, March 1999, *67* (2), 251–333.

Acemoglu, D. and J. Angrist, "How Large are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," *NBER Macroeconomics Annual 2000*, 2006.

Albanese, Andrea and Giovanni Gallo, "Buy flexible, pay more: The role of temporary contracts on wage inequality," *Labour Economics*, June 2020, *64*, 101814.

Anders, Jenna and Charlie Rafkin, "The Welfare Effects of Eligibility Expansions: Theory and Evidence from SNAP," June 2022.

Andrews, I. and T. B. Armstrong, "Unbiased Instrumental Variables Estimation Under Known First-Stage Sign," *Quantitative Economics*, 2017.

Andrews, M.J., L. Gill, T. Schank, and R. Upward, "High wage workers match with high wage firms: Clear evidence of the effects of limited mobility bias," *Economics Letters*, December 2012, *117* (3), 824–827.

Angrist, J. and A. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 1991.

_ **, G. Imbens, and E. Rubin**, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996.

Araujo, Ana Luisa Pessoa De, "Wage Inequality and Job Stability," preprint, Institute Working Paper December 2017.

Armstrong, Tim and Michal Kolesár, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018.

_ **and** _ , "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 2021.

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska, "Concentration in US Labor Markets: Evidence From Online Vacancy Data," *Labour Economics*, July 2020, p. 101886.

Babet, Damien, Olivier Godechot, and Marco G Palladino, "In the Land of AKM: Explaining the Dynamics of Wage Inequality in France," 2022, p. 62.

Bassier, Ihsaan, Arindrajit Dube, and Suresh Naidu, "Monopsony in Movers The Elasticity of Labor Supply to Firm Wage Policies," *Journal of Human Resources*, April 2022, *57* (S), S50–s86. Publisher: University of Wisconsin Press.

Baumberg, Ben, "The stigma of claiming benefits: a quantitative study," *Journal of Social Policy*, April 2016, *45* (2), 181–199. Publisher: Cambridge University Press.

Behaghel, Luc, B. Crépon, and Marc Gurgand, "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment," *American Economic Journal: Applied Economics*, 2014.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen, "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain.," *Econometrica*, 2012.

Bentolila, Samuel, Juan J. Dolado, Wolfgang Franz, and Christopher Pissarides, "Labour Flexibility and Wages: Lessons from Spain," *Economic Policy*, April 1994, *9* (18), 53.

Besley, Timothy and Stephen Coate, "Understanding welfare stigma: Taxpayer resentment and statistical discrimination," *Journal of Public Economics*, July 1992, *48* (2), 165–183.

Bhargava, Saurabh and Dayanand Manoli, "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment," *American Economic Review*, November 2015, *105* (11), 3489–3529.

Blanchard, Olivier and Augustin Landier, "The Perverse Effects of Partial Labour Market Reform: Fixed-Term Contracts in France," *The Economic Journal*, 2002, *112* (480), F214–F244. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0297.00047.

Bodner, Ronit and Drazen Prelec, "Self-signaling and diagnostic utility in everyday decision making," in "Collected Essays in Psychology and Economics," Oxford University Press, 2002.

Bonhomme, Stéphane and Grégory Jolivet, "The pervasive absence of compensating differentials," *Journal of Applied Econometrics*, 2009, *24* (5), 763–795. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.1074.

Bursztyn, Leonardo and Robert Jensen, "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure," *Annual Review of Economics*, 2017, *9* (1), 131–153. _eprint: https://doi.org/10.1146/annurev-economics-063016-103625.

_ , Georgy Egorov, and Robert Jensen, "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education," *The Review of Economic Studies*, July 2019, *86* (4), 1487–1526.

_ , _ , Ruben Enikolopov, and Maria Petrova, "Social Media and Xenophobia: Evidence from Russia," Working Paper 26567, National Bureau of Economic Research December 2019. Series: Working Paper Series.

Bénabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, December 2006, *96* (5), 1652–1678.

_ , Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Reasoning," July 2018.

Caggese, Andrea and Vicente Cuñat, "Financing Constraints and Fixed-Term Employment Contracts," *The Economic Journal*, 2008, *118* (533), 2013–2046. Publisher: [Royal Economic Society, Wiley].

Cahuc, Pierre, Olivier Charlot, and Franck Malherbet, "Explaining the Spread of Temporary Jobs and Its Impact on Labor Turnover," *International Economic Review*, 2016, *57* (2), 533–572. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12167.

_ , _ , _ , Helène Benghalem, and Emeline Limon, "Taxation of Temporary Jobs: Good Intentions with Bad Outcomes?," *The Economic Journal*, February 2020, *130* (626), 422–445.

Card, David, Ana Rute Cardoso, and Patrick Kline, "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *," *The Quarterly Journal of Economics*, May 2016, *131* (2), 633–686.

Cardoso, Ana Rute, Joerg Heining, Patrick Kline, and David Card, "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 2018, *36*, 58.

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng, "On Binscatter," October 2022. arXiv:1902.09608 [econ, stat].

Celhay, Pablo A., Bruce D. Meyer, and Nikolas Mittag, "Errors in Reporting and Imputation of Government Benefits and Their Implications," Working Paper 29184, National Bureau of Economic Research August 2021. Series: Working Paper Series.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," *arXiv*, 2021.

Claeskens, G and N Hjort, "The Focused Information Criterion," *Journal of the American Statistical Association*, 2003.

Coussens, Stephen and Jann Spiess, "Instrumental Variable Estimation with First-Stage Heterogeneity," *Working Paper*, 2021.

Créchet, Jonathan, "Risk Sharing in a Dual Labor Market," 2023.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté, "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco," *American Economic Journal: Applied Economics*, 2015.

Daruich, Diego, Sabrina Di Addario, and Raffaele Saggio, "The Effects of Partial Employment Protection Reforms: Evidence from Italy," *The Review of Economic Studies*, February 2023, p. rdad012.

DellaVigna, Stefano and Ethan Kaplan, "The Fox News Effect: Media Bias and Voting," *Quarterly Journal of Economics*, August 2007, *122* (3), 49.

Dellavigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao, "Voting to Tell Others," *The Review of Economic Studies*, January 2017, *84* (1), 143–181.

DellaVigna, Stefano, Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya, "Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia," *American Economic Journal: Applied Economics*, July 2014, *6* (3), 103–132.

Deshpande, Manasi and Yue Li, "Who Is Screened Out? Application Costs and the Targeting of Disability Programs," *American Economic Journal: Economic Policy*, November 2019, *11* (4), 213–248.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)" Association for Computational Linguistics Minneapolis, Minnesota June 2019, pp. 4171–4186.

Djourelova, Milena, "Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration," *American Economic Review*, March 2023, *113* (3), 800–835.

Donoho, D. L., "Statistical estimation and optimal recovery," *Annals of Statistics*, 1996.

Drenik, Andres, Simon Jäger, Pascuel Plotkin, and Benjamin Schoefer, "Paying outsourced labor: Direct evidence from linked temp agency-worker-client data," *THE REVIEW OF ECONOMICS AND STATISTICS*, 2022, p. 11.

Dube, Arindrajit, Alan Manning, and Suresh Naidu, "Monopsony and Employer Misoptimization Explain Why Wages Bunch at Round Numbers," Technical Report w24991, National Bureau of Economic Research, Cambridge, MA September 2018.

_ , Suresh Naidu, and Adam Reich, "Power and Dignity in the Low-Wage Labor Market: Theory and Evidence from Wal-Mart Workers," Technical Report w30441, National Bureau of Economic Research, Cambridge, MA September 2022.

Dubois, Hans, Anna Ludwinek, and European Foundation for the Improvement of Living and Working Conditions, eds, *Access to social benefits: reducing non-take-up* number 15/36. In 'EF.', Luxembourg: Publications Office of the European Commission [u.a.], 2015. OCLC: 928849932.

Dynarski, Susan, "The Economics of Student Aid," 2007.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya, "Media and Political Persuasion: Evidence from Russia," *American Economic Review*, December 2011, *101* (7), 3253–3285.

Finkelstein, Amy and Matthew J Notowidigdo, "Take-Up and Targeting: Experimental Evidence from SNAP*," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1505–1556.

Foos, Florian and Daniel Bischof, "Tabloid Media Campaigns and Public Opinion: Quasi-Experimental Evidence on Euroscepticism in England," *American Political Science Review*, August 2021, pp. 19–37. Publisher: Cambridge University Press.

Franceschin, Riccardo, "Choosing Employment Protection: the role of On-the-Job Search and Ability Learning," 2023.

Friedrichsen, Jana, Tobias König, and Renke Schmacker, "Social image concerns and welfare take-up," *Journal of Public Economics*, December 2018, *168*, 174–192.

Frölich, Markus, "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 2007, *139* (1), 35 – 75. Endogeneity, instruments and identification.

Gavin, Neil T., "Below the radar: A U.K. benefit fraud media coverage tsunami—Impact, ideology, and society," *The British Journal of Sociology*, 2021, *72* (3), 707–724. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-4446.12809.

Geiger, Ben Baumberg, "Disabled but not deserving? The perceived deservingness of disability welfare benefit claimants," *Journal of European Social Policy*, March 2021, p. 0958928721996652. Publisher: SAGE Publications Ltd.

_ , Kate Bell, and Declan Gaffney, "Benefits stigma in Britain," Project report, Turn2us, Elizabeth Finn Care November 2012.

Ghosal, Sayantan, Smarajit Jana, Anandi Mani, Sandip Mitra, and Sanchari Roy, "Sex Workers, Stigma, and Self-Image: Evidence from Kolkata Brothels," *The Review of Economics and Statistics*, May 2022, *104* (3), 431–448.

Goffman, Erving, *Stigma : notes on the management of spoiled identity*, New York, N.Y.: Simon & Schuster, 1986.

Güell, Maia, "Fixed-Term Contracts and Unemployment: An Efficiency Wage Analysis," *SSRN Electronic Journal*, 2000.

_ **and José Vicente Rodríguez**, "TEMPORARY CONTRACTS, INCENTIVES AND UNEMPLOYMENT," 2010, p. 36.

Hansen, C. and D. Kozbur, "Instrumental variables estimation with many weak instruments using regularized JIVE.," *Journal of Econometrics*, 2012.

Holford, Angus, "Take-up of Free School Meals: Price Effects and Peer Effects," *Economica*, 2015, *82* (328), 976–993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecca.12147.

Homonoff, Tatiana and Jason Somerville, "Program Recertification Costs: Evidence from SNAP," *American Economic Journal: Economic Policy*, November 2021, *13* (4), 271–298.

Hong, Han and Denis Nekipelov, "Semiparametric efficiency in nonlinear LATE models," *Working paper*, 2010.

Huntington-Klein, Nick, "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," *Journal of Causal Inference*, 2020.

Imbens, G. and J. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics," *The Economic Journal*, 2017.

Ivandic, Ria, Tom Kirchmaier, and Stephen J. Machin, "Jihadi Attacks, Media and Local Hate Crime," *SSRN Electronic Journal*, 2019.

Jarosch, Gregor, "Searching for Job Security and the Consequences of Job Loss," *Econometrica*, 2023, *91* (3), 903–942. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14008.

Jäger, Simon, Christopher Roth, U Cologne, Nina Roussille, and Benjamin Schoefer, "Worker Beliefs About Outside Options," Technical Report March 2023.

Kitagawa, T and C. Muris, "Model averaging in semiparametric estimation of treatment effects," *Journal of Econometrics*, 2016.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar, "Who Profits from Patents? Rent-Sharing at Innovative Firms," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1343–1404. Publisher: Oxford Academic.

_ , Raffaele Saggio, and Mikkel Sølvsten, "Leave-Out Estimation of Variance Components," *Econometrica*, 2020, *88* (5), 1859–1898.

Lagrosa, Ivan, "Income Dynamics in Dual Labor Markets," September 2022.

Lavetti, Kurt and Ian M Schmutte, "Estimating Compensating Wage Differentials with Endogenous Job Mobility," 2018, p. 68.

Leeb, H. and B. Pötscher, "Model selection and inference-Facts and Fiction," *Econometric theory*, 2005.

Levy, Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, March 2021, *111* (3), 831–870.

Link, Bruce G. and Jo C. Phelan, "Conceptualizing Stigma," *Annual Review of Sociology*, August 2001, *27* (1), 363–385.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," July 2019. arXiv:1907.11692 [cs].

Major, Brenda and Laurie T. O'Brien, "The Social Psychology of Stigma," *Annual Review of Psychology*, February 2005, *56* (1), 393–421.

Manning, Alan, "Monopsony in Labor Markets: A Review," *ILR Review*, January 2021, *74* (1), 3–26.

Marie, Etienne and Vincent Jaouen, "Evaluation du contrat à durée déterminée dit d'usage (CDDU)," IGAS rapport December 2015.

Marinescu, Ioana, Ivan Ouss, and Louis-Daniel Pape, "Wages, hires, and labor market concentration," *Journal of Economic Behavior & Organization*, April 2021, *184*, 506–605.

Matsaganis, Manos, Nirina Rabemiafara, and Terry Ward, "Young people and temporary employment in Europe," Eurofound report January 2014.

Mijovic-Prelec, Danica and Drazen Prelec, "Self-deception as self-signalling: a model and experimental evidence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, January 2010, *365* (1538), 227–240. Publisher: Royal Society.

Morales, Juan S., "Legislating during war: Conflict and politics in Colombia," *Journal of Public Economics*, January 2021, *193*, 104325.

Morrison, James, *Scroungers: Moral Panics and Media Myths*, Zed Books, 2019.

Müller, Karsten and Carlo Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime," *Journal of the European Economic Association*, August 2021, *19* (4), 2131–2167.

_ **and** _ , "From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment," *American Economic Journal: Applied Economics*, 2023.

Narayan, Ayushi, "Does simplifying the college financial aid process matter?," *Economics of Education Review*, April 2020, *75*, 101959.

Oreopoulos, P., "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, 2006.

Osman, Adam and Jamin D Speer, "Stigma and Take-Up of Labor Market Assistance: Evidence from Three Experiments," *Economica*, 2023.

Pescosolido, Bernice A. and Jack K. Martin, "The Stigma Complex," *Annual Review of Sociology*, August 2015, *41* (1), 87–116.

Postel–Vinay, Fabien and Jean–Marc Robin, "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity," *Econometrica*, 2002, *70* (6), 2295–2350. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2002.00441.x.

Rion, Normann, "Waiting for the Prince Charming: Fixed-Term Contracts as Stopgaps," 2021, p. 52.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," February 2020. arXiv:1910.01108 [cs].

Singh, Rahul and Liyang Sun, "Automatic Kappa Weighting for Instrumental Variable Models of Complier Treatment Effects," *Working paper*, 2021.

Sorkin, Isaac, "Ranking Firms Using Revealed Preference*," *The Quarterly Journal of Economics*, August 2018, *133* (3), 1331–1393.

Staiger, E. and J. Stock, "Instrumental variables regressions with weak instruments," *Econometrica*, 1997.

Stephens, M. and D. Yang, "Compulsory Education and the Benefits of Schooling," *American Economic Review*, 2014.

Stuber, Jennifer and Karl Kronebusch, "Stigma and other determinants of participation in TANF and Medicaid," *Journal of Policy Analysis and Management*, 2004, *23* (3), 509–530. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pam.20024.

Sunstein, Cass R., "Nudges.gov: Behavioral Economics and Regulation," February 2013.

Słoczyński, T., "When Should We (Not) Interpret Linear IV Estimands as LATE?," *Working Paper*, 2022.

Vytlacil, E., "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 2002.

List of Tables

List of Figures