Improving LATE estimation in experiments with imperfect compliance

Yagan Hazard (**Job Market Paper**) Simon Löwe*

December 1, 2022

[Please click here for the latest version]

Abstract

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.

Keywords: IV, LATE, imperfect compliance, precision, variance

JEL Codes: C01, C21, C26

*Hazard: PhD candidate, Paris School of Economics. Löwe: PhD candidate, Paris School of Economics. We would like to thank warmly Luc Behaghel and Xavier D'Haultfœuille for their invaluable and unfailing support — without which this research would have been undoubtedly far less enjoyable and rewarding. Y. Hazard is deeply grateful to Toru Kitagawa, Jon Roth, Peter Hull, Soonwoo Kwon, Emily Oster and Jesse Shapiro for enlightening discussions at Brown University. We also thank Philipp Ketz, Marc Gurgand, Eric Maurin, and seminar participants at the PSE-CREST internal seminar and the Brown Econometrics Coffee for helpful comments and discussions. All remaining errors are our own.

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 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 "naïvely" 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

[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.

[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.

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.

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 $\sigma^2_{"}$) is homoscedastic.[3] Denoting by N the sample size, p the share of encouraged individuals, and π the share of compliers, we get:[4]

$$\text{Var}\left[\widehat{LATE}^{2SLS}\right] = \frac{1}{N} \quad \frac{1}{\pi^2} \quad \frac{\sigma^2_{"}}{p \quad (1 \quad p)}$$

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

[3] Here, " 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 might seem similar. Instead, p is the share of individuals who are incentivized (or assigned) to take the treatment, while 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:

$$V^{2SLS} = \frac{2}{\varepsilon}\left[\quad DZ \quad \frac{1}{ZZ} \quad ZD\right] \quad ^1 = \frac{\frac{2}{\varepsilon}}{(E[DjZ = 1] \quad E[DjZ = 0])^2 \quad E[Z](1 \quad E[Z])} = \frac{\frac{2}{\varepsilon}}{^2 \quad p(1 \quad p)}$$

where we used the notation $_{XY}$ $E[XY^0]$ and the definitions $= E[DjZ = 1] \quad E[DjZ = 0]$ and $p = E[Z]$.

times larger to reach the same precision (variance) as in the perfect compliance experiment. Put it 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.

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 in 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]

[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

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 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.

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 affected by the

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.

[8] An estimator $\hat{\theta}$ of a parameter θ has a "first-order" or "asymptotic" bias when the limiting distribution of $\sqrt{n}\left(\hat{\theta} - \theta\right)$ is not centered on 0. For instance, if $\hat{\theta}$ is asymptotically normal with first-order bias B, then we have: $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(B, 1)$. Notice that it does not prevent $\hat{\theta}$ from being a consistent estimator of θ. Yet it indicates that it does not converges 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: $\mathbb{E}(\hat{\theta}) - \theta = O(1/\sqrt{n})$. When used in this way, $\hat{\theta}$ can still be asymptotically biased while $\sqrt{n}$-rate consistent.

randomly modified incentive.[9]

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.[10] 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 be targeted by policy-makers by using the exact same encouragement device (instrument) as in the experimental setting. Ul-

[9] 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 birth date. 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 birth date and the associated compulsory schooling laws.

[10] 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.

timately, 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.[11] 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.[12] 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.[13]

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.

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 po-

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

[12] 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.

[13] 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.

tential 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).[14] We have:

$$Y = D \quad Y(1) + (1 \quad D) \quad Y(0)$$
$$D = Z \quad D(1) + (1 \quad Z) \quad D(0)$$

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 $f(Y_i, D_i, Z_i, G_i)g_{1;\dots;n}$ 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:* $(Y(1), Y(0), D(1), D(0), G)$? Z[15]
2. *Exclusion restriction:* $Y(D, Z) = Y(D)$
3. *First Stage:* $\mathrm{E}[D = 1 \; j \; Z = 1] \quad \mathrm{E}[D = 1 \; j \; Z = 0] > 0$
4. *Monotonicity:* $D(1) \quad D(0)$

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.[16] 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 effect among compliers $\mathrm{E}[Y(1) \quad Y(0) j D(1) > D(0)]$, is identified. The usual estimator for the LATE is the Wald estimator — which coincides with the two-stage-least-squares (2SLS) estimator in our

[14] 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.

[15] 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.

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

case where D and Z are binary:

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

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:

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

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[17] (or no compliers at all) is what creates room for precision gains in the estimation of the (overall) LATE.[18] 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 none of

[17] 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).

[18] 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:

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

the compliers are women:

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

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.

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

where $\Delta_n^{W|G=g} \quad \mathrm{E}_n[W|Z = 1, G = g] \quad \mathrm{E}_n[W|Z = 0, G = g]$. 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:

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

where $\sigma_"^2$ denotes the variance of the errors,[19] 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 decreases the variance by a factor $(1 \quad \Pr[G = 0])$. This is intuitive: the more we can get rid of

[19] Here, " 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: " = $Y \quad \text{LATE} \; D$.

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.[20]

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. □

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 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

[20] 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 > 0 (or < 0) instead of $\neq 0$ (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.

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 a additional first-order bias.[21]

Simulations presented in appendix C tend to confirm such concerns. We report in Table 1 below the results of a Monte-Carlo simulation using DGP0 described in appendix C. 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 — 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 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).

Given the issues documented with the "naïve" approach presented above, we propose an 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).

[21] 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.

Estimator 2 Test-and-Select

1: Divide the sample in two equally sized random sub-samples I_1 and I_2 — stratifying the random split by G.

2: Within subsample 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 I_2 the groups for which we rejected — in sample 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 I_2.

5: Repeat steps 2 to 4 reversing the roles of I_1 and I_2 (cross-fitting).

6: Take the average of the estimators obtained in step 4 within I_1 and 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. The remaining bias despite the use of data splitting and cross-fitting could be explained by the finite sample bias of 2SLS estimator.[22]

Table 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.

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

[22] 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. 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.

using data-splitting.

3 THEORETICAL RESULTS

Throughout this section, we will consider a framework with two i.i.d. samples: a *test* sample (denoted I_T) used in order to t-test on group-specific first-stage coefficients, and an *estimation* sample (denoted I_E) used in order to compute the resulting estimator with the selection rule induced by the tests' results in 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 $p_E = 1 \quad p_T$) going to sample I_T (respectively I_E). We will denote by n_T (= p_T n) and n_E (= p_E n) the corresponding sample sizes — and we will use the notation $n \; ! \; 1$ 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 I_T and 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.[23] 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:

[23] 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.

1. $G_S = \{$all groups with strong first stage$\}$

2. $G_W = \{$all groups with weak first stage$\}$

3. $G_0 = \{$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:* $G_W = \emptyset$.

Let $S \in \{0,1\}^{|G|}$ 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:

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

In words, $\hat{\tau}(S)$ is the Wald estimator on the subsample such that $S_{G_i} = 1$, which is the subsample designated by S. As an example, for $|G| = 2$,

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

$$= \begin{cases} \hat{\tau}^{WALD} & \text{if} \quad S_1 = S_0 = 1 \\ \hat{\tau}_1^{WALD} & \text{if} \quad S_1 = 1 \quad \& \quad S_0 = 0 \\ \hat{\tau}_0^{WALD} & \text{if} \quad S_1 = 0 \quad \& \quad S_0 = 1 \end{cases} \tag{2}$$

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 I_T — constructed as a random split of the initial sample.[24] 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 I_T. We can then define:

$$\hat{\tau}_E = \hat{\tau}\left(\hat{S}_{(T)}\right) \tag{3}$$

which is ultimately our estimator of interest, the TS estimator computed on split I_E.[25] Equivalently, for any selection vector S, we wille denote by $\hat{\tau}_E(S)$ the estimator computed on the subsample defined by S, within split 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* $E[|Y|^2] < 1$*, 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 B.

□

Notice that it would be possible to decrease the threshold of the t-test at an appropriate rate 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 S_{strong} this subset of all selection vectors (i.e., a subset of $\{0,1\}^{|G|}$) that never excludes groups with strong first-stages. Formally, for any $\tilde{S} \in S_{strong}$, we

[24] This vector stacks the $|G|$ test decisions resulting from our $|G|$ t-tests (one per group) in I_T.

[25] 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.

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

Proof. See appendix A. □

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:

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

where $\hat{V}_E(S)$ 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.:

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

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 to rescale CIs by $\sqrt{n_E}$ to allow for a meaningful comparison.[26]

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*

[26] 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).

estimator computed in split I_E) are such that:

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

Moreover, we have that:

$$\mathrm{P}\left[LATE \in CI\ (\hat{S})\right]_{n \to \infty} \geq 1 \quad \alpha$$

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

Proof. See appendix A.

□

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 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.[27] 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,

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

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.[28] Then instead of dividing by two the size of the estimation sample (post-selection), she ends up reducing it by a factor of $\frac{4}{5} \quad \frac{1}{2} = \frac{2}{5} < \frac{1}{2}$ — 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 $\frac{5}{2} \quad \frac{1}{4} = \frac{5}{8} > \frac{1}{2}$. 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 to make the testing-selecting and estimation steps independent. As shown in lemma 1 and illustrated in Table 1, our procedure would yield a biased estimator in the absence of sample splitting.

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, I_1 and I_2. She constructs a first estimator using I_1 as the test sample and I_2 as the estimation sample, and a second using I_2 as the test sample and 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

[28] 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.

ones discussed above for the sample split estimator. Hence averaging those estimators would potentially yield an estimator with the same variance than 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 I_1 and I_2 are asymptotically independent one from another.*

Proof. See appendix A. □

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:

$$V\left(\frac{N(0,\hat{V_1}) + N(0,\hat{V_2})}{2}\right) = \frac{\hat{V_1} + \hat{V_2}}{4} = \frac{\hat{V_1}}{2}$$

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 $\frac{p}{2}$ 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, overall the gain in asymptotic variance described in the above display exactly compensate 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.[29] 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 asymp-

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

totically 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:*

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

All such values of g for which first-stages are weak are gathered in G_W.

In parallel, the data-generating process is assumed to be such that for any group g, the share of observations contained in the group is constant (it does not vary with n), nor does the LATE within the group. Formally:

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

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

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

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 $E[|Y|^2] < +\infty$, 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 B. □

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: $\forall g \in G_S,\ S_g = 1$ (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 $E[|Y|^2] < +\infty$, we have:*

1. $\forall S \in S_{strong}$,[30] $\sqrt{n_E}\,(\hat{\tau}(S) - LATE) \xrightarrow{d} \mathcal{N}(B(S), V^S)$.
2. $B(S) \neq 0 \Rightarrow |LATE^{W(S)} - LATE| > 0$, *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.* $\{S_j = 0 \setminus j \in G_W\}$ *and* $LATE^{W(S)} \neq LATE$.

Proof. See appendix A. □

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

[30] Recall that S_{strong} is defined such that for any $S \in S_{strong}$, we have: $\forall g \in G_S;\ S_g = 1$.

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:

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

where $\pi \quad \Pr[D(1) > D(0)]$ is the share of compliers in the population, $LATE^g \quad E[Y(1) - Y(0) \mid D(1) > D(0),\ G = g]$ 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 $\sqrt{\bar{n}}$-rate under assumption 3. Hence even once rescaled by $\sqrt{\bar{n}}$, the bias (with respect to the LATE parameter) remains bounded as long as the treatment effect heterogeneity term $\left(LATE^1 \quad LATE^2\right)$ 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, $E[|Y|^2] < \infty$, and homogeneous treatment effects, we have that $\hat{\tau}(S)$ is first-order unbiased and asymptotically normal, i.e.:*

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

Less restrictively, under assumptions 1 and 3, $E[|Y|^2] < \infty$, and vanishing treatment effect heterogeneity, i.e.:

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

$\hat{\tau}(S)$ *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.[31]

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* $E[|Y|^2] < \infty$*, 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* $$\sqrt{n_E}\frac{\hat{\tau}_E - LATE}{\sqrt{\hat{V}_E}} \xrightarrow{d} \mathcal{N}(0,1)$$

[31] 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.

Proof. See appendix A. □

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 B for a proof of the bias of Coussens and Spiess (2021) procedure under our framework.[32] 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 (unsquared). 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

[32] 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.

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 (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 κ% 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:

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

where we assumed $\mathrm{Var}[Y(0)] = \mathrm{Var}[Y(1)] = \sigma^2$,[33] and q_x is the x^{th}-quantile of a $N(0,1)$. 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

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

of the study designed to detect them. In such a case, assumption 4 would be fulfilled.

4 EXTENSIONS

In this section, we present some of the extensions we plan to develop in future versions of this paper.

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.

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:

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

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.[34] The quantity $\mathrm{P}[G = 2 | D(1) > D(0)]$ can be estimated by 2SLS as suggested in Abadie (2003).[35] 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 asymptotically valid $(1 \quad \alpha)$-confidence interval around this parameter — e.g., using inversion of an Anderson-Rubin statistic. Suppose we construct 99%-CI

[34] 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.

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

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 $(LATE^2 \quad LATE^1) \quad M$, 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 $M \ \widehat{UB}^P$. Suppose we widen our 96%-CI around $LATE^1$ — the effect among compliers in the population selected by our procedure — by $M \ \widehat{UB}^P$. Such CI is (asymptotically) valid with a 95% coverage for the overall LATE parameter.[36] 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 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.

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 $s(X) \quad \mathrm{E}[D(1) \quad D(0) | X]$, 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:
   $\mathrm{E}[s(X) | \hat{s}(X) \ 2 \ [q_{g \ 1}, q_g]] = \mathrm{E}[D(1) \quad D(0) | \hat{s}(X) \ 2 \ [q_{g \ 1}, q_g]] \quad \pi^g$

That way, we are back to a situation in which the covariates are reduced to a partition of the population: $f \mathbb{1}_{\hat{s}(x) 2 [q_{g \ 1}; q_g]} g_{g 2 f 1; ...; G g}$. Unfortunately, the procedure proposed in Chernozhukov

[36] 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 0:96 = 0:9504.

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,[37] 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 $\Pr[\text{sh. of compliers} = 0|g \in G_S]$ 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).

[37] 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 =2 to construct p-values and CIs of level . 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.

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 ($\mathrm{E}[Y(1) \quad Y(0) | D(1) > D(0),\ G = g]$) and compliance rates ($\mathrm{E}[D(1) \quad D(0) | G = g]$).

DGP parameters To simulate a flexible DGP, we use the threshold crossing model representation (Vytlacil, 2002).[38] Let $(\delta_i, \varepsilon_i)' \quad N\ (0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} \sigma \ = 1 & \rho \ '' \\ \rho \ '' & \sigma'' = 1 \end{pmatrix}$$

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 $\rho \ ''$ the correlation coefficient between δ_i and ε_i. The potential treatment indicators are given by:

$$D_i(0) = \mathbb{1}(\Phi \ (\delta_i) < S_{AT}), \qquad D_i(1) = \mathbb{1}(\Phi \ (\delta_i) < 1 \quad S_{NT})$$

where Φ denotes the cdf of a $N\left(\vec{0}, \Sigma\right)$, 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:

$$D_i = D_i(0) \ (1 \quad Z_i) + D_i(1) \ Z_i$$

We also define a covariate X as:

$$X_i = \delta_i + \eta_i$$

[38] 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.

where η_i $N(0, \sigma^2)$. The covariate X is therefore a noisy predictor of treatment receipt. We also define groups G as the J-quantiles of X:

$$G_i = \mathbb{1}\left(F(X_i)\ 2\ \left[\frac{j\quad 1}{J}, \frac{j}{J}\right]\right)$$

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

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

where $\tilde{\pi}_G = \pi_G\quad E_G(\pi_G)$ is the centered compliance rate by group with $G(i)$ representing the group G of individual i and ν_i $N(0, \sigma^2{}_G)$. 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:

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

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,[39] 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:

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

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

4. ρ_{ν}: correlation between latency to treat and baseline untreated potential. Controls selection into treatment and hence the necessity for instrumentation.

5. σ^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:

$$\text{DGP1} \quad \left(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\nu} = 0.3, \sigma = 0.01, \alpha = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\} \right)$$

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

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

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. We vary treatment effect heterogeneity, and quantify the latter on the x-axis by scaling the standard deviation of $Y(1) \quad Y(0)$ by the Minimum Detectable Effect (MDE):

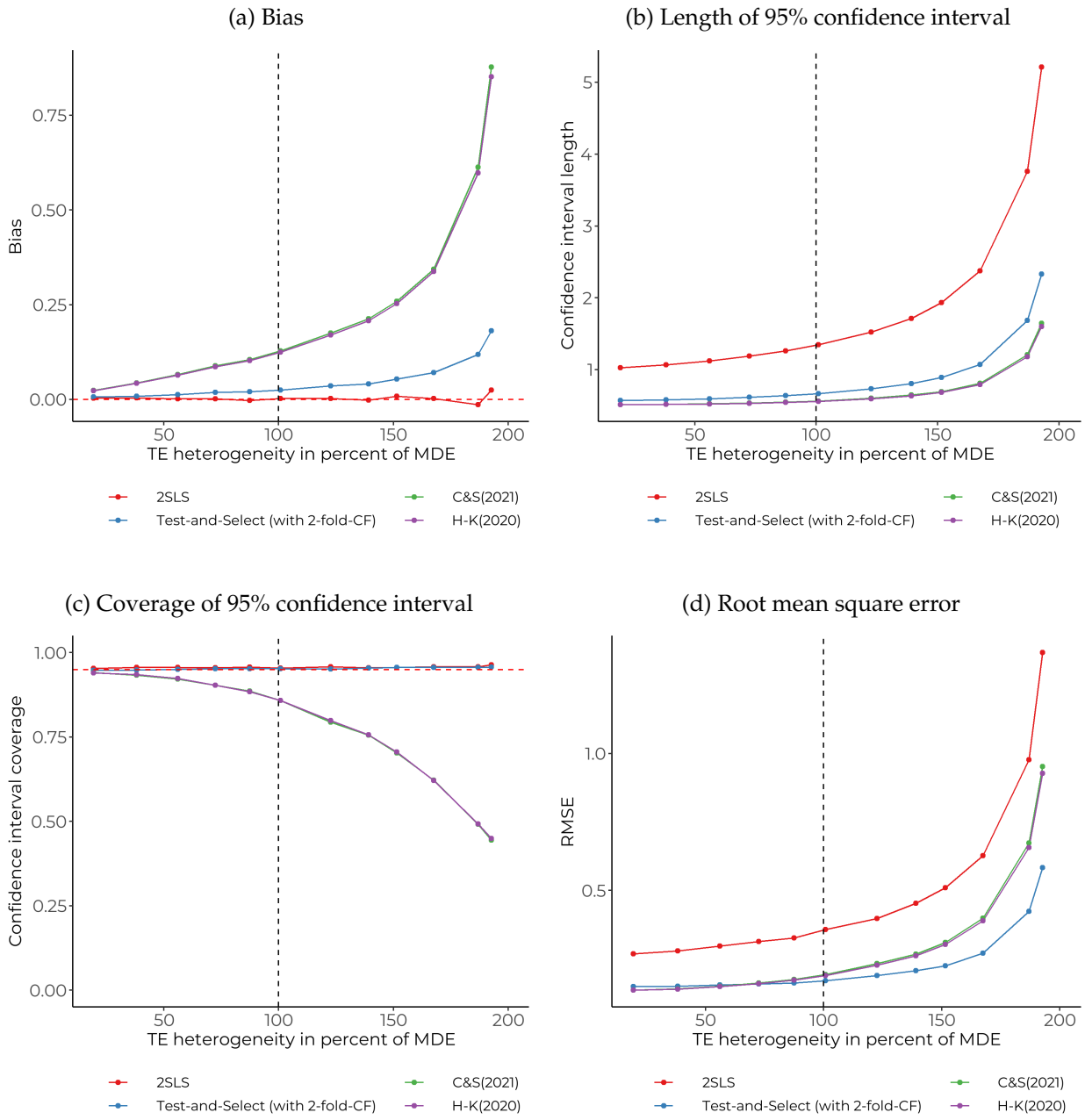
$$x = \frac{\sqrt{V(Y(1) \quad Y(0))}}{\sqrt{\frac{V(Y(1)) + V(Y(0))}{0.5\ n}} \quad \frac{}{q_{0.975} + q_{0.8}}}$$

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 in 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 presents the bias, length and coverage of 95%-CIs, and RMSE of the different estimators considered in these simulations under DGP1. Panel 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 (ex-

cept zero), as expected. This translates in poor coverage rates of these estimation strategies, when ours covers at the nominal level for any amount of treatment effect heterogeneity — see panel 1c. Moreover, panel 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 very similar gains in precision compared to Huntington-Klein (2020) or Coussens and Spiess (2021)'s estimators, hence the similar reduction in CI length. Overall, this leads logically to a domination of our method in terms of RMSE in such a setting — see panel 1d. Notice that in general, Huntington-Klein (2020) or Coussens and Spiess (2021)'s estimators could very well display a larger decrease in variance compared to our methodology. Yet one of the points illustrated in this simulation is that such gain would come along with some bias as long as treatment effect heterogeneity and (conditional) compliance rates are correlated, thus possibly hurting badly inference on the LATE.

Figure 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:

$$\text{DGP2} \quad \left(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho \text{ " } = 0.3, \sigma \ = 0.5, \alpha = 0.5, \beta \ 2 \ f1, 2, 3, 4, 6, 10, 20, 40, 80g\right)$$

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

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

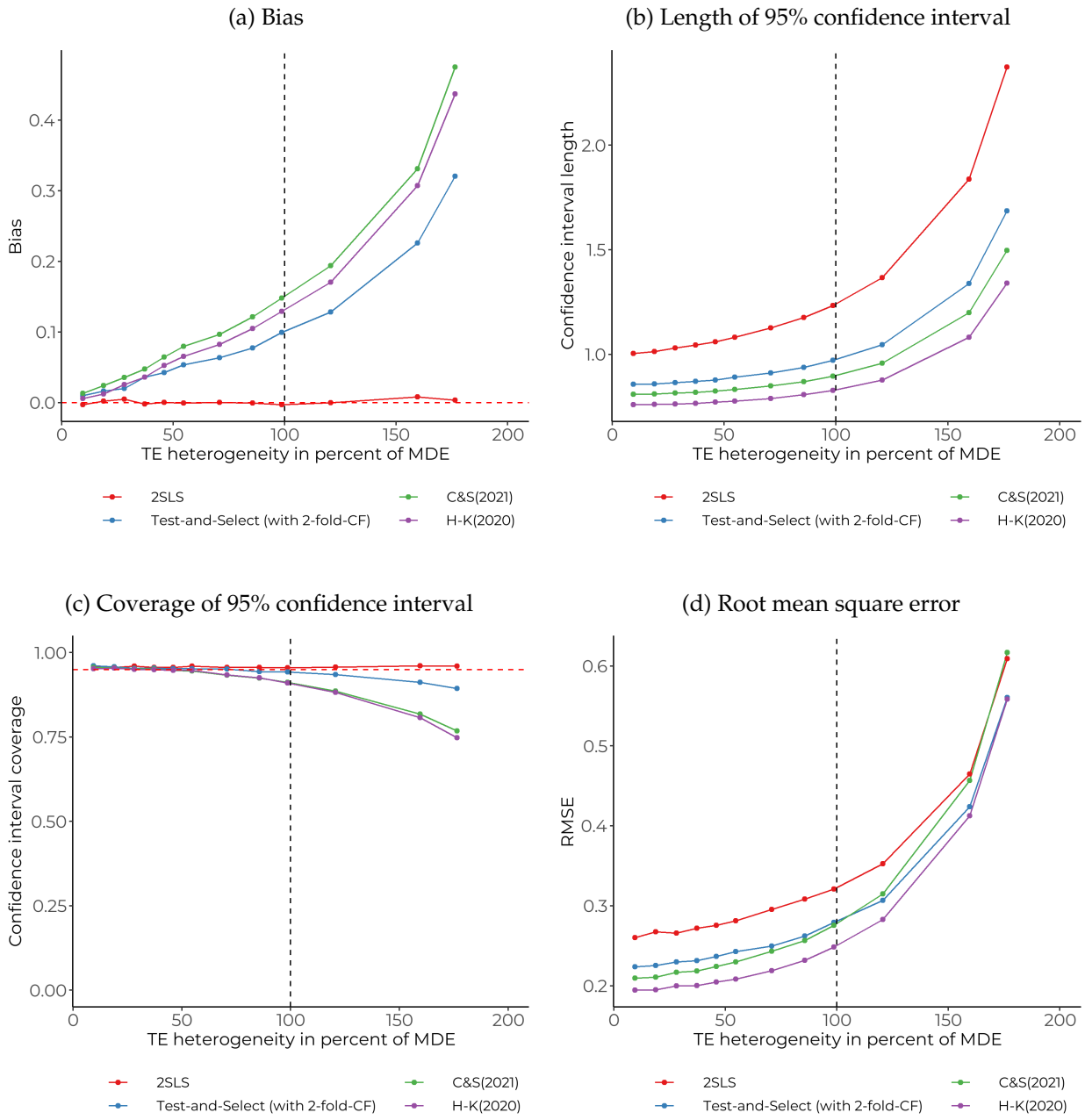
for an overall compliance of 25%, as in DGP1. In the spirit of varying parameters as little as possible across DGPs, we keep the same $\alpha = 0.5$ as in DGP1, which again means that there is a significant correlation between compliance and treatment effects across groups.

We run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 2. Compared to what we observed in DGP1 — and as expected from our theoretical results from section 3 — panel 2a highlights a much larger bias of our procedure, that grows as treatment effect heterogeneity increases. Therefore, at first glance we could expect very similar performance of our estimator compared to the ones proposed in Huntington-Klein (2020) and Coussens and Spiess (2021) as their level of bias seems relatively similar. Yet panel 2a conceals the different *distributions* of such estimators compared to ours. Indeed, and as implicitly illustrated in panel 2b, Huntington-Klein (2020) and Coussens and Spiess (2021)'s estimators have a lower variance than ours, yielding significantly shorter 95%-CIs. In absence of any bias, this would unequivocally be synonymous of a better performance. Yet since neither our procedure nor theirs is unbiased, these shorter CIs yield worse coverage properties than our estimator — see panel 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 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 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 $x = 100$, while this is not the case of alternative estimators — except for the standard 2SLS of course. Lastly, panel 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.[40]

[40] 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 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).[41] Compulsory schooling laws (that restrict the age at which individuals are allowed to drop out of school) vary across states and along 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 along 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 to make 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 sex) US census division 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.[42] Indeed, we do not want to identify the effect of compulsory schooling laws on education by 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.

[41] 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.

[42] 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.

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 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 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 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 sex) US census division 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).

Table 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 3 reports the results of estimators that control (somehow linearly) for 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 in a linear way 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). An alternative is to use another estimator than the 2SLS estimator to 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):

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

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 4. One can observe that the resulting point estimates differ quite a lot from the ones previously documented in table 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 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. 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. 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 sex) birth state 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 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. 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 sex) birth state 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.[43] 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 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

[43] 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.

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 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 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 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

Table 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.

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 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 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 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 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 selection rule based on a 0.99 level t-test. The last but one column corresponds to a 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 behaviour 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 exploiting first-stage heterogeneity in a different way 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.

REFERENCES

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231 – 263, 2003. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(02)00201-4. URL http://www.sciencedirect.com/science/article/pii/S0304407602002014.

Alberto Abadie, Jiaying Gu, and Shu Shen. Instrumental variable estimation with first-stage heterogeneity. *Working paper*, 2022.

D. Acemoglu and J. Angrist. How large are human-capital externalities? evidence from compulsory schooling laws. *NBER Macroeconomics Annual 2000*, 2006.

I. Andrews and T. B. Armstrong. Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, 2017.

J. Angrist and A. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 1991.

J. Angrist, G. Imbens, and E. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 1996.

Tim Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 2018.

Tim Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 2021.

Luc Behaghel, 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.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 2012.

K. Borusyak and P. Hull. Non-random exposure to exogenous shocks. *Working Paper*, 2021.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv*, 2021.

G Claeskens and N Hjort. The focused information criterion. *Journal of the American Statistical Association*, 2003.

Stephen Coussens and Jann Spiess. Instrumental variable estimation with first-stage heterogeneity. *Working Paper*, 2021.

Bruno Crépon, 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.

D. L. Donoho. Statistical estimation and optimal recovery. *Annals of Statistics*, 1996.

Markus Frölich. Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35 – 75, 2007. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2006.06.004. URL http://www.sciencedirect.com/science/article/pii/S0304407606001023. Endogeneity, instruments and identification.

C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 2012.

Han Hong and Denis Nekipelov. Semiparametric efficiency in nonlinear late models. *Working paper*, 2010.

Nick Huntington-Klein. Instruments with heterogeneous effects: Bias, monotonicity, and localness. *Journal of Causal Inference*, 2020.

G. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 1994.

John P. A. Ioannidis, T. D. Stanley, and Hristos Doucouliagos. The power of bias in economics. *The Economic Journal*, 2017.

Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects, 2018.

T Kitagawa and C. Muris. Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics*, 2016.

H. Leeb and B. Pötscher. Model selection and inference-facts and fiction. *Econometric theory*, 2005.

P. Oreopoulos. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, 2006.

Rahul Singh and Liyang Sun. Automatic kappa weighting for instrumental variable models of complier treatment effects. *Working paper*, 2021.

E. Staiger and J. Stock. Instrumental variables regressions with weak instruments. *Econometrica*, 1997.

M. Stephens and D. Yang. Compulsory education and the benefits of schooling. *American Economic Review*, 2014.

T. Słoczyński. When should we (not) interpret linear iv estimands as late? *Working Paper*, 2022.

E. Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 2002.

A 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:

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

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 $(1 \quad \alpha)$ — 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:

$$\{\text{Group G=0 is selected}\} \ , \ \left\{ \mathbb{1}\left\{ \sqrt{n^0} \ \frac{\hat{\pi}^0}{\hat{\sigma}^{\wedge 0}} > q_1 \right\} = 1 \right\}$$

With probability tending to $(1 \quad \alpha)$, 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 $LATE_1$:

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

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:

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

where $\hat{P}_g \quad \hat{\mathrm{P}}[G = g] = n_g/n$ and $\widehat{ITT}_g$ denotes the difference-in-means estimator of the intention-to-treat estimand ($E[Y|Z = 1] \quad E[Y|Z = 0]$) 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 $\sqrt{n}\ \hat{\pi}^0 \xrightarrow[n \to \infty]{d} N(0, V_{\hat{\pi}^0})$. 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 $\sqrt{n}\ \hat{\pi}^0$ being larger than a given threshold t. We trivially have:

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

where $N(0, LB = t, V_{\hat{\pi}^0})$ denote the distribution of a truncated normal distribution $N(0, V_{\hat{\pi}^0})$ 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 $ITT^0 \xrightarrow{p} 0$ and, by Slutsky's lemma $\sqrt{n}\,\hat{\pi}^0\, ITT^0 \xrightarrow{d} \mathcal{N}(0, \tilde{V}_0)$. Yet once we condition on the estimated first-stage coefficient, the probability limit of ITT^0 and the limiting distribution of $\sqrt{n}\,\hat{\pi}^0\, ITT^0$ are quite different. Indeed, we have:

$$ITT^0 \mid \hat{\pi}^0 = f \xrightarrow{p} f \quad B_{AT} \quad _{NT}$$

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

$$\sqrt{n}\,\hat{\pi}^0\, ITT^0 \mid \sqrt{n}\,\hat{\pi}^0 > t \xrightarrow{d} \mathcal{N}(0, LB = t, V_{\wedge 0}) \quad B_{AT} \quad _{NT}$$

If $B_{AT} \quad _{NT} = 0$, then this limiting distribution becomes degenerate at 0, and the second bias term B is null. Yet if $B_{AT} \quad _{NT} = 0$, then this additional term B is not centered at 0, and therefore it adds an additional first-order bias to the estimator $LATE$. 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 ($B_{AT} \quad _{NT} = 0$), 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 ($B_{AT} \quad _{NT} \neq 0$), 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 $S \in \{0,1\}^{|G|}$. S_{G_i} is merely the G_i^{th} line of the vector S. Let us also us the following notation:

- $G_S = \{\text{all groups with strong fist stage}\}$
- $G_0 = \{\text{all groups with zero fist stage}\}$

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 $S \in S_{\text{strong}}$. 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. $\{g \in G_S, S_g = 1\} \setminus \{g \in G_0, S_g = 0\}$
2. $\{g \in G_S, S_g = 1\} \setminus \{g \in G_0, S_g = 1\}$

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

$$\hat{A} = \left(\sum_i Z_i S_{G_i}\right)^{-1} \sum_i Z_i S_{G_i} Y_i, \qquad A = E[Y|Z=1, S_G=1]$$

$$\hat{B} = \left(\sum_i ((1 \quad Z_i) S_{G_i})\right)^{-1} \sum_i (1 \quad Z_i)\, S_{G_i} Y_i, \qquad B = E[Y|Z=0, S_G=1]$$

$$\hat{C} = \left(\sum_i Z_i S_{G_i}\right)^{-1} \sum_i Z_i S_{G_i} D_i, \qquad C = E[D|Z=1, S_G=1]$$

$$\hat{D} = \left(\sum_i ((1 \quad Z_i) S_{G_i})\right)^{-1} \sum_i (1 \quad Z_i)\, S_{G_i} D_i, \qquad D = E[D|Z=0, S_G=1]$$

$$\Rightarrow LATE = \frac{A \quad B}{C \quad D}$$

$$\Rightarrow \hat{\tau}(S) = \frac{\hat{A} \quad \hat{B}}{\hat{C} \quad \hat{D}}$$

Notice that the fact that $LATE = \frac{A \quad B}{C \quad D}$ 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:

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

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

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

Therefore we get:

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

where $\varepsilon \quad Y \quad LATE \quad D$ is the structural error term of the second stage, and $p_{C;S_G=1} = \mathrm{E}[D(1) \quad D(0) | S_G = 1]$ is the share of compliers among the selected. As expected from an influence function,

one can check that $E[\psi_{\hat{(S)};i}] = 0$. It follows that asymptotically,

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

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

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

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

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

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

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

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

Therefore, we only need to prove that:

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

This is proven below:

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

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 $V^{\wedge(S)}$, 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 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:

$$
\hat{S}_{g;(T)} \quad 1\left\{ \hat{\pi}^g_{(T)} > \frac{\hat{\sigma}^{\ g}}{\sqrt{n^g_{(T)}}} \quad q_1 \right\}
$$

where $n^g_{(T)}$ is the number of observations in group g in sample I_T, $\hat{\pi}^g_{(T)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\ g}$ 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:

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

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

$$\sqrt{n_{(E)}}(\hat{\tau}_T \quad LATE) = \frac{1}{\sqrt{n_{(E)}}} \sum_i \psi_{\wedge_T;i}$$

where the expression of the influence function is given by:

$$\psi_{\wedge_T;i} = \frac{1}{p_{C;\hat{S}_{G,T}=1}} \left(\frac{1}{E[Z\hat{S}_{G;T}]} Z_i \hat{S}_{G_i;T} \quad (\varepsilon_i \quad E[\varepsilon | Z=1, \hat{S}_{G;T}=1]) \right.$$
$$\left. \frac{1}{E[(1 \quad Z)\hat{S}_{G;T}]} (1 \quad Z_i)\hat{S}_{G_i;T} \quad (\varepsilon_i \quad E[\varepsilon | Z=0, \hat{S}_{G;T}=1]) \right)$$

The above display makes it clear that the $\psi_{\wedge_T;i}$'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 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 $\psi_{\wedge_T;i}$'s, and we are back to the case studied in proposition 1.1 and 1.2. Now let us define:

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

where $V^{\wedge_E}$ 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:

$$\mathrm{E}[e^{it\hat{T}_E} | \hat{S}_T] = \sum_{S \in \{0;1\}^{|G|}} \mathbf{1}\{\hat{S} = S\} \quad \mathrm{E}[e^{it\hat{T}_E} | \hat{S}_T = S]$$
$$\frac{p}{n!} \frac{!}{1} \sum_{S \in S_{strong}} \mathbf{1}\{\hat{S} = S\} \quad e^{\quad t^2/2} + \sum_{S \notin S_{strong}} 0 \quad \mathrm{E}[e^{it\hat{T}_E} | \hat{S}_T = S]$$

Indeed, by proposition 1.1 we have that for $\hat{S}_T \in S_{strong}$, $\hat{T}_E$ 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: $\forall S \in S_{strong}, \ \mathbf{1}\{\hat{S}_T = S\} \quad \frac{p}{n!} \frac{!}{1} \quad 0$, hence the second line of the above display (by continuous mapping theorem).

Notice that by Jensen inequality: $|\mathrm{E}[e^{it\hat{T}_E} | \hat{S}_T]| \quad \mathrm{E}[|e^{it\hat{T}_E}| | \hat{S}_T] = 1$, hence by the dominated con-

vergence theorem we get:

$$\mathrm{E}[e^{it\hat{T}_E}] = \mathrm{E}\left[\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T]\right] \xrightarrow{p} E\left[\sum_{S \in S_{strong}} \mathbf{1}\{\hat{S} = S\} \; e^{-t^2/2} + \sum_{S \notin S_{strong}} 0\right]$$
$$= e^{-t^2/2} \quad \text{(characteristic function of a } N(0,1)\text{)}$$

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 C) imply convergence in 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:

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

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

□

Proof. **Corollary 1.**

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

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

where $\hat{V}(S)$ 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 $N(0,1)$ distribution. Those CIs are asymptotically valid by proposition 1.1, i.e.:

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

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

$$\mathrm{P}[LATE\ 2\ CI\ \ (\hat{S})] = E\left[E[\mathbf{1}\{LATE\ 2\ CI\ \ (\hat{S})\}|\hat{S} = S]\right] \underset{n\to\infty}{\to} 1 \quad \alpha$$

which is the second statement of corollary 1.

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

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

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

$$V^{\wedge_E(S)} < V^{\mathrm{TSLS}}$$

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

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

and

$$\forall\varepsilon_2 > 0,\ \lim_{n\to\infty} \mathrm{P}[\left|\sqrt{n_E}\ \ \mathrm{length}[CI^{TSLS}] \quad \sqrt{n_E}\ \ \mathrm{length}[CI^{0;TSLS}]\right| > \varepsilon] = 0$$

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

$$\sqrt{n_E}\ \ \mathrm{length}[CI^0(S) < \sqrt{n_E}\ \ \mathrm{length}[CI^{0;TSLS}]$$

Hence we have:

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

□

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 I_2 as the test sample and I_1 as the estimation sample, recall that we can decompose it as follows:

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

where the expression of the influence function is given by:

$$\psi_{\hat{}_1;i} = \frac{1}{p_{C;\hat{S}_{G,(2)}=1}} \left(\frac{1}{E[Z\hat{S}_{G;(2)}]} Z_i \hat{S}_{G_i;2} \quad (\varepsilon_i \quad E[\varepsilon | Z=1, \hat{S}_{G;(2)}=1]) \right.$$
$$\left. \frac{1}{E[(1 \quad Z)\hat{S}_{G;(2)}]} (1 \quad Z_i)\hat{S}_{G_i;2} \quad (\varepsilon_i \quad E[\varepsilon | Z=0, \hat{S}_{G;(2)}=1]) \right)$$

with $\hat{S}_{g;(2)}$ denoting the selection indicator for group g computed in fold I_2 as follows:

$$\hat{S}_{g;(2)} \quad 1\left\{ \hat{\pi}^g_{(2)} > \frac{\hat{\sigma}^{\ g}}{\sqrt{n^g_{(2)}}} \quad q_1 \right\}$$

where $n^g_{(2)}$ is the number of observations in group g in sample I_1, $\hat{\pi}^g_{(2)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\ g}$ is a consistent estimator of the (asymptotic) variance of $\hat{\pi}^g_{(2)}$. Second, recall (from the proof of corollary 1 above) that:

$$\sqrt{n^g_{(2)}} \left[\hat{\pi}^g_{(2)} \quad \pi g \right] = \frac{1}{\sqrt{n^g_{(2)}}} \left[\sum_{i=1}^{n^g_{(2)}} \underbrace{\left(\frac{Z_i (D_i \quad \mathsf{E}[D \mid Z=1])}{E[Z]} + \frac{(1 \quad Z_i) \quad (D_i \quad \mathsf{E}[D \mid Z=0])}{1 \quad \mathsf{E}[2]} \right)}_{\tilde{}^g_i} \right]$$
$$= \frac{1}{\sqrt{n^g_{(2)}}} \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i$$

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 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 $\mathrm{P}[\hat{S}_{g;(2)} = 1] \underset{n \to \infty}{\to} 1$ 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 I_2), we have:

$$\begin{aligned}
\hat{S}^{(n^g_{(2)})}_{g;(2)} &= \mathbb{1}\left\{ \hat{\pi}^g_{(2)} > \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}}} q_1 \right\} \\
&= \mathbb{1}\left\{ \frac{1}{\sqrt{n^g_{(2)}}} \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i > \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}}} q_1 \right\} \\
&= \mathbb{1}\left\{ F^{g;n^g_{(2)}} > \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}}} q_1 \right\}
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g;(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g;(2)} = 1\right] \\
&= \mathrm{P}\left[F^{g;n^g_{(2)}-1} \leq \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}-1}} (q_1 - \epsilon), \quad F^{g;n^g_{(2)}} > \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}}} q_1 \right] \\
&\quad \mathrm{P}\left[\left|F^{g;n^g_{(2)}-1}\right| \leq \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}-1}} (q_1 - \epsilon), \quad \left|F^{g;n^g_{(2)}}\right| > \frac{\hat{\sigma}^g}{\sqrt{n^g_{(2)}}} q_1 \right] \\
&= \mathrm{P}\left[\left| \left(n^g_{(2)} - 1\right) F^{g;n^g_{(2)}-1} \right| \leq \sqrt{n^g_{(2)}-1}\, \hat{\sigma}^g (q_1 - \epsilon), \quad n^g_{(2)} \left|F^{g;n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\, \hat{\sigma}^g q_1 \right]
\end{aligned}$$

Notice that:

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

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

$$
\begin{aligned}
&\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g;(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g;(2)} = 1\right] \\
&\quad \mathrm{P}\left[\left|\left(n^g_{(2)}-1\right) F^{g;n^g_{(2)}-1}\right| \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\ (q_1-\epsilon),\ n^g_{(2)}\left|F^{g;n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ q_1\right] \\
&= \mathrm{P}\left[\left|\left(n^g_{(2)}-1\right) F^{g;n^g_{(2)}-1}\right| \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\ (q_1-\epsilon),\ n^g_{(2)}\left|F^{g;n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ q_1\right] \\
&\quad \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ q_1 \left(n^g_{(2)}-1\right)\left|F^{g;n^g_{(2)}-1}\right|, \left(n^g_{(2)}-1\right)\left|F^{g;n^g_{(2)}-1}\right| \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ (q_1-\epsilon)\right] \\
&\quad \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ q_1 \left(n^g_{(2)}-1\right)\left|F^{g;n^g_{(2)}-1}\right|, \right. \\
&\qquad \left. \left(n^g_{(2)}-1\right)\left|F^{g;n^g_{(2)}-1}\right| \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\ (q_1-\epsilon)\right] \\
&\quad \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \sqrt{n^g_{(2)}}\ \hat{\sigma}^{g}\ q_1 \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\ (q_1-\epsilon)\right] \\
&= \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \hat{\sigma}^{g}\ q_1 \left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\ \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\right]
\end{aligned}
$$

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

$$\hat{\sigma}^{g}\ q_1 \left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\ \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g} \quad \epsilon\ \sqrt{n^g_{(2)}-1}\ \sigma^{g}\ _{n\,!\,1}\,!\ 1$$

Hence we get:

$$\mathrm{P}\left[\hat{S}^{(n^g_{(2)}-1)}_{g;(2)} = 0,\ \hat{S}^{(n^g_{(2)})}_{g;(2)} = 1\right] \quad \mathrm{P}\left[\left|\tilde{\psi}^g_{n^g_{(2)}}\right| > \hat{\sigma}^{g}\ q_1 \left(\sqrt{n^g_{(2)}}-\sqrt{n^g_{(2)}-1}\right) + \epsilon\ \sqrt{n^g_{(2)}-1}\ \hat{\sigma}^{g}\right]\ _{n\,!\,1}\,!\ 0$$

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 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 S_{strong}. When S takes its value in some subsets of S_{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:

$$S^0_{\text{strong}} \quad \{S \in S_{\text{strong}} : \forall g \in G_W, S_g = 1\}$$
$$S^1_{\text{strong}} \quad \{S \in S_{\text{strong}} : \exists g \in G_W, S_g = 0\}$$

By proposition 1 and the argument above, we have:

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

Now, we turn to the case where S belongs to the set S^1_{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 S^1_{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 S_{strong} — that is, all the values that the random vector $\hat{S}$ (determined in sample I_T) takes with

non-zero probability asymptotically — the asymptotic bias of $\sqrt{n_E}\,(\hat{\tau}(S) \quad LATE)$ is of the form:

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

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: $LATE^{\mathcal{G}_W^S} \quad LATE^{\mathcal{G}_S} = o(1)$. 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:

$$V^{\wedge(S)} \quad V^{TSLS}$$

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.

□

B PROOFS OF USEFUL LEMMAS

Proof. **Lemma 2.**

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

$$T^g_{;n_T} = \mathbb{1}\left\{ \sqrt{n^g_T} \quad \frac{\hat{\pi}^g}{\hat{\sigma}^g} > q_1 \quad \right\}$$

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 $q_1 \quad \frac{\alpha}{2}$ the $1 \quad \overline{2}$ quantile of a $N(0,1)$.

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

$$\forall g \in G_S, \quad \lim_{n_T \to \infty} \Pr[T^g_{;n_T} = 1] = 1$$

since we have: $\forall g \in G_S,\ \pi^g > 0$.

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

$$\forall g \in G_0, \quad \lim_{n_T \to \infty} \Pr[T^g_{;n_T} = 1] = \alpha$$

since we have: $\forall g \in G_0,\ \pi^g = 0$. □

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:

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

The quantiles of a $|N(b,1)|$ 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:

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

□

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:

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

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

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

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:

$\frac{\hat{A}}{\hat{B}} - \frac{A}{B} = o_P(1)$.

□

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)**.**

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

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

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

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:

$$\hat{A} = \left(\sum_i Z_i\right)^{1} \sum_i Z_i Y_i, \qquad A = E[Y|Z=1]$$

$$\hat{B} = \left(\sum_i (1 \quad Z_i)\right)^{1} \sum_i (1 \quad Z_i)\, Y_i, \qquad B = E[Y|Z=0]$$

$$\hat{C} = \left(\sum_i Z_i\right)^{1} \sum_i Z_i D_i, \qquad C = E[D|Z=1]$$

$$\hat{D} = \left(\sum_i (1 \quad Z_i)\right)^{1} \sum_i (1 \quad Z_i)\, D_i, \qquad D = E[D|Z=0]$$

$$) \; LATE = \frac{A \quad B}{C \quad D}$$

$$) \; \hat{\tau}^{Wald} = \frac{\hat{A} \quad \hat{B}}{\hat{C} \quad \hat{D}}$$

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

$$a_i = \frac{Z_i(Y_i \quad E[Y|Z=1])}{E[Z]}$$

$$b_i = \frac{(1 \quad Z_i)(Y_i \quad E[Y|Z=0])}{1 \quad E[Z]}$$

$$c_i = \frac{Z_i(D_i \quad E[D|Z=1])}{E[Z]}$$

$$d_i = \frac{(1 \quad Z_i)(Y_i \quad E[D|Z=0])}{1 \quad E[Z]}$$

We then have:

$$\sqrt{n}(\hat{\tau}^{Wald} \quad LATE) = \frac{1}{\sqrt{n}} \sum_i \psi_{Wald;i} + o_P(1)$$

where (following lemma 5) $\psi_{\sim Wald;i}$ is given by:

$$\begin{aligned}
\psi_{\sim Wald;i} &= \frac{(a_i \quad b_i) \quad LATE \quad (c_i \quad d_i)}{C \quad D} \\
&= \frac{1}{\pi}\left(\frac{Z_i(Y_i \quad E[YjZ=1])}{E[Z]} \quad \frac{(1 \quad Z_i)(Y_i \quad E[YjZ=0])}{1 \quad E[Z]}\right. \\
&\qquad\qquad \left. LATE \quad \left(\frac{Z_i(D_i \quad E[DjZ=1])}{E[Z]} \quad \frac{(1 \quad Z_i)(D_i \quad E[DjZ=0])}{1 \quad E[Z]}\right)\right) \\
&= \frac{1}{\pi}\left(\frac{1}{p}Z_i \quad (\varepsilon_i \quad E(\varepsilon jZ=1)) \quad \frac{1}{1 \quad p}(1 \quad Z_i) \quad (\varepsilon_i \quad E(\varepsilon jZ=0))\right)
\end{aligned}$$

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

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

where $V(\psi_{\wedge Wald;i})$ equals:

$$\begin{aligned}
V(\psi_{\wedge Wald;i}) &= E(\psi^2_{\wedge Wald;i}) \\
&= E(\psi^2_{\wedge Wald;i} jZ=1)p + E(\psi^2_{\wedge Wald;i} jZ=0)(1 \quad p) \\
&= \frac{1}{\pi^2}\left(\frac{1}{p}E[(\varepsilon \quad E[\varepsilon jZ=1])^2 jZ=1] + \frac{1}{1 \quad p}E[(\varepsilon \quad E[\varepsilon jZ=0])^2 jZ=0]\right) \\
&= \frac{1}{\pi^2}\left(\frac{1}{p}V[\varepsilon jZ=1] + \frac{1}{1 \quad p}V[\varepsilon jZ=0]\right)
\end{aligned}$$

□

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:*

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

with $B(S) = \frac{H^2 \Pr[G=2]}{} \quad (LATE^1 \quad LATE^2)$ *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 ($\pi^2 = H^2/\sqrt{\overline{n}}$), 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 ($\sqrt{\overline{n}}$ 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[44] computed on the whole estimation sample... hence it is $\sqrt{\overline{n}}$ consistent (no asymptotic bias).
In the second case, the resulting estimator corresponds to the Wald estimator computed on group 1. Hence it is a $\sqrt{\overline{n}}$ consistent estimator for the LATE conditional on being in group 1, which we define below:

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

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

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

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

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

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

[44] 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).

We have:

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

where the last line uses our standard notations:

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

Hence we get:

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

Therefore, we have in this case:

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

with $B(S^2) \quad \frac{H^2 \ \Pr[G=2]}{} \ \big(LATE^1 \quad LATE^2\big)$.

□

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: $\tau(X) = \lambda + \frac{\zeta(X)}{\sqrt{n}}$.

Their proof goes as follows:

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

where:

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

The convergence of $\sqrt{n}(\hat{\tau}_w - \tau_w)$ 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:

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

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

Indeed, we have:

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

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

C 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:

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

We vary the number of groups as a way to exacerbate the pre-testing issue in simulations. We report in Table 1 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 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.[45]

Table 7: Pre-test bias, and the use of cross-fitting

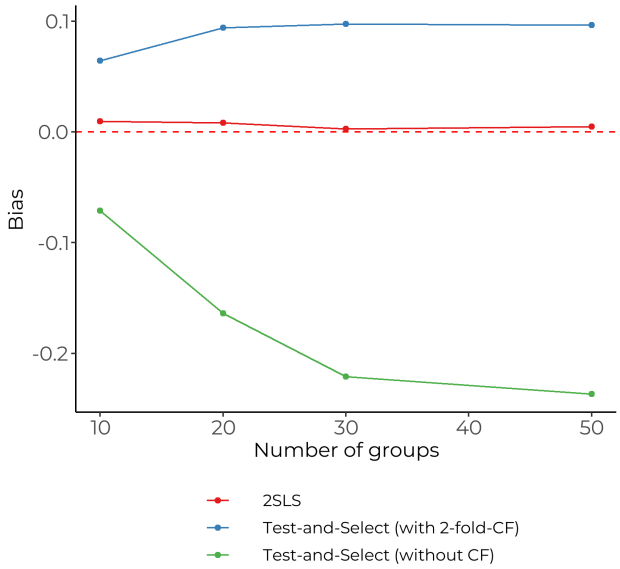
	2SLS	Test-and-Select (with 2-fold-CF)	Test-and-select (without CF)
Bias	0.003	0.097	-0.221
Coverage	0.953	0.976	0.861

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 3 reports the bias of the three estimators presented in Table 1 for a varying number of

[45] 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. 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.

Figure 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.

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.