Improving LATE estimation
in experiments with imperfect compliance

Yagan Hazard   Simon Löwe*

September 30, 2022

[Please click here for the latest version]

Abstract

Experimental evaluations of many policies of interest (e.g., educational and training programs) inevitably face incomplete treatment group take-up. Estimation of causal effects in these "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 clarify under which conditions 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 is illustrated using Monte Carlo simulations and an application to a large-scale experiment on job search counselling.

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

JEL Codes: C01, C21, C26

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. 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 their precision by taking advantage of such heterogeneity in their estimation. 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 (where possible) IV estimation to sub-populations with non-zero compliance rates. 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.[1]

We first underline the pitfalls of "naïvely" implementing such a selection rule based on estimated compliance rates, and propose a simple fix to this issue by resorting to data-splitting. Next, we study the asymptotic properties of the Test-and-Select estimator under both standard and weak-IV-like asymptotic sequences. The former analysis allows us to illustrate the potential gains in precision while the latter aims at better approximating the finite sample properties of our proposed estimator. These analyses underline the robustness of the Test-and-Select procedure to

---

[1]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.

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 an application to a large-scale experiment on job search counseling. These sections illustrate the improved robustness of our estimator to treatment effect heterogeneity.

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 simplifying case where treatment effects are homogeneous and the variance of the errors (denoted σ_ε^2)[2] is homoscedastic. Denoting by N the sample size, p the share of encouraged individuals, and π the share of compliers, we get:

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

Here the importance of the compliance rate appears clearly. Let's take an illustrative example, studying the variance in two experiments evaluating the same program, one with a 10% compliance rate ($\pi^C = 0.1$) and another with a perfect compliance ($\pi^C = 1$). The compliance rate in the first experiment is only 10 times lower than in the second experiment, and yet the sample size needs to be a 100 times larger to reach the same precision (variance) as in the perfect compliance experiment. Therefore, 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, leading to possibly uninformative inference.[3]

Another way to approach the example above is the following. Suppose researchers from the first experiment were able to identify the sub-population of compliers — based on some observables. Then focusing on this fraction (10%) of the population would reduce the estimation sample

---

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

[3]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.

by 10, but it would still *decrease* variance by a factor of 10 and therefore greatly improve inference without changing the overall sample size. In this paper, we propose a two-step procedure that searches for and excludes from the estimation sample sub-populations that do not respond to the instrument — i.e., sub-populations without compliers. To fix ideas in a more concrete setting, consider the same-sex instrument (Angrist and Evans, 1998), which builds on the idea that some groups of parents have preferences for gender diversity among their children, leading them to have a third child if and only if the first two are of the same sex. Fertility preferences have been documented to be vastly heterogeneous and depend on multiple dimensions (religion, opportunity cost of child bearing etc.). Therefore, some sub-populations might not react to the same-sex instrument, and as such they would not contribute to the identification of the LATE. The existence of such non-compliant groups is not a threat to identification.[4] However, we make it clear that their presence in the estimation sample does reduce the precision with which the LATE is estimated. As a simple and intuitive fix, we propose a Test-and-Select estimator that drops groups without compliers from the estimation sample.

Under standard asymptotics, 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,[5] 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. 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

---

[4]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.

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

is such that it would not be systematically detected in finite samples. We discuss in detail why this is a relevant 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 randomly modified incentive.[6]

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

---

[6]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.

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 re-visited the IV estimation strategy to try to achieve precision gains. This renewed interest can be related to empirical attempts to identify treatment effects "on those who take it up" (Crépon et al., 2015). More specifically, the paper most closely related to our work is Coussens and Spiess (2021). They propose to use the IV estimator that is optimal in the constant treatment effect regime which, in the more general case, 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 het-erogeneity. 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.[7] 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. Ultimately, choosing between our approach and one in the vein of Huntington-Klein (2020) or Coussens and Spiess (2021) 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 pop-ulation have a share of compliers well separated from 0.[8] 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.[9] Our work also uses a two-step procedure akin to the one studied in Abadie et al. (2022). However, we focus on the more general setting of heterogeneous treatment effects, and only consider the case of a binary treatment and binary instrument.[10]

---

[7]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.

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

[9]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.

[10]Abadie et al. (2022) do propose an interpretation of the estimand targeted by their methodology under heteroge-neous 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.

The remainder of the paper is structured 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 an empirical application to 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 $(Y(1), Y(0), D(1), D(0), Z, X)$, where $(Y(1), Y(0)$ are the potential outcomes when treated or not, $(D(1), D(0)$ are the potential outcomes when encouraged or not, Z is the encouragement status, and G is a discrete predetermined covariate (assumed binary in this section for illustrative purposes).[11] We have:

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

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

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

ASSUMPTION 1 (LATE identifying assumptions).

  1. *Independence:* $(Y(1), Y(0), D(1), D(0), G) \perp Z$[12]

  2. *Exclusion restriction:* $Y(D, Z) = Y(D)$

  3. *First Stage:* $\mathrm{E}\left[D = 1 \mid Z = 1\right] - \mathrm{E}\left[D = 1 \mid Z = 0\right] > 0$

---

[11]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.

[12]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.

4. Monotonicity: $D(1) \geq D(0)$

Under those assumptions, Angrist et al. (1996) showed that the LATE, defined as the average treatment effect among compliers $E[Y(1) - Y(0)|D(1) > D(0)]$, is identified. The usual estimator for the LATE is the Wald estimator — which coincides with the two-stage-least-squares (2SLS) estimator in our case where D and Z are binary:

$$\widehat{\text{LATE}}^{2SLS} = \frac{\left(\sum_i Z_i\right)^{-1} \sum_i Z_i Y_i - \left(\sum_i (1 - Z_i)\right)^{-1} \sum_i (1 - Z_i) Y_i}{\left(\sum_i Z_i\right)^{-1} \sum_i Z_i D_i - \left(\sum_i (1 - Z_i)\right)^{-1} \sum_i (1 - Z_i) D_i}$$
$$= \frac{E_n[Y|Z = 1] - E_n[Y|Z = 0]}{E_n[D|Z = 1] - E_n[D|Z = 0]}$$

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 realization of Z and D. To fix ideas, let us think of G as a sex indicator taking value 0 for men, 1 for women. We allow for heterogeneous shares of compliers across sex, i.e., women might react more (or less) than men to the encouragement. Formally:

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

We do not impose that both π^0 and π^1 are strictly larger than 0, only that the average share of compliers in the population is well separated from 0 (assumption 1.3). In other words, we allow for one of the two groups to be absolutely fully unresponsive to the encouragement — as long as the other is, allowing the identification of the overall LATE. This is key in our reasoning, as considering the existence of sub-populations with very few[13] compliers (or no compliers at all) is what creates room for precision gains in the estimation of the (overall) LATE.[14] 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

---

[13]This vague terminology ("very 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).

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

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

case, women's observations do not bring any signal in the estimation of the overall LATE, as none of the compliers are women:

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

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

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

where $\Delta_n^{W|G=g} \equiv \text{E}_n[W|Z = 1, G = g] - \text{E}_n[W|Z = 0, G = g]$. 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 ($V^{TSLS,G=1}$) with the one of a 2SLS estimator on the full sample (V^{TSLS}), assuming homoscedasticity of the errors and homogeneous treatment effects:

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

where σ_ε^2 denotes the variance of the errors [15], N the sample size and p the share of encouraged

---

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

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

Motivated by this illustrative example, we propose the following Test-and-Select (TS) estimation procedure:

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).[16]

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 first-stage coefficient.

LEMMA 1 (Pre-testing and first-stage first-order bias). *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 will lead to a first-order bias in the estimation of the first-stage coefficient.*

Proof. See appendix A. □

Such results suggest that there might be significant distortions due to the pre-testing step of the suggested procedure that might yield non-valid inference ultimately. Simulations presented in appendix C tend to confirm such concerns. We report in Table 4 below the results of a Monte-Carlo

---

[16]Usually, in the context of RCTs, researchers will have a strong prior about the way their encouragement affects the treatment status, hence the ability to use as an alternative hypothesis $\pi > 0$ (or $\pi < 0$) instead of $\pi \neq 0$. 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.

simulation using DGP0 described in section 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 — 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 4 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.[17]

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 im-

---

[17]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 4. 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.

plement the selection of groups with or without compliers in a given sample. In section 3 and as already introduced above, we propose to use data-splitting to fix the pre-testing issues previously mentioned, and we suggest the use of cross-fitting to alleviate the efficiency loss incurred when using data-splitting.[18]

3 THEORETICAL RESULTS

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

The study of the properties of our suggested estimator will be divided into two parts. Firstly, we will consider the case where covariate-defined sub-groups contain either a share of compliers well-separated from zero, or no compliers at all. This case will simplify the study of the potential precision gains derived from the suggested procedure. In a second step, we will introduce groups with a "local-to-zero" (or "weak") share of compliers, à la Staiger and Stock (1997) — meaning that the share of compliers in those groups decreases at a $1/\sqrt{n}$ rate, placing them in the same order of magnitude as sampling variation. Such a modeling choice is made in an effort to better approximate the finite-sample behavior of the estimator, by allowing for imperfect selection of groups with non-zero shares of compliers.[19] 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 $G = g$.

---

[18]Future versions of this paper will further investigate under which conditions cross-fitting fully compensates for the potential efficiency loss due to sample splitting.

[19]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.

We denote by $\mathcal{G}$ the support of G. In order to distinguish groups with "strong", "weak" and zero shares of compliers, we will further define:

1. $\mathcal{G}_S = \{\text{all groups with strong first stage}\}$

2. $\mathcal{G}_W = \{\text{all groups with weak first stage}\}$

3. $\mathcal{G}_0 = \{\text{all groups with zero first stage}\}$

3.1 Standard asymptotics

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

ASSUMPTION 2 (No weak first-stages). *There are no groups for which the share of compliers is local-to-zero. Formally:*

$$\forall g, \ g \in \mathcal{G}_S \cup \mathcal{G}_0$$

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

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

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 $|\mathcal{G}| = 2$,

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

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

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 $\mathcal{I}_T$ — constructed as a random split of the initial sample.[20] We will denote the latter by $\hat{S}^{(T)}$, where the hat and superscript (T) indicate that this vector comes from an estimation step in sample $\mathcal{I}_T$. We can then define:

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

which is ultimately our estimator of interest.

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

LEMMA 2 (Asymptotic distribution of the selection procedure). *Under assumptions 1 and 2, 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 $\hat{\tau}(S)$ when both test and estimation sample size tends to

---

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

infinity, we can restrict ourselves to vectors S which select at least all vectors with strong first stages. We denote by $\mathcal{S}_{strong}$ this subset of all selection vectors (i.e., a subset of $\{0,1\}^{|\mathcal{G}|}$) that never excludes groups with strong first-stages. Formally, for any $\tilde{S} \in \mathcal{S}_{strong}$, we have: $\forall g \in \mathcal{G}_S,\ \tilde{S}_g = 1$.

PROPOSITION 1. *Under assumptions 1 and 2, we have:*

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

2. $\forall S \in \mathcal{S}_{strong},\ V(\hat{\tau}(S)) \leq V^{TSLS} \quad$ *with equality iff:* $\quad \forall g, \quad S_g = 1$ *or in degenerate cases*

3. $\sqrt{n_E} \cdot \dfrac{\hat{\tau}_T - LATE}{\sqrt{\hat{V}(\tau(\hat{S}_T))}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$

Proof. See appendix A. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

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

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

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

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

The following corollary states that such intervals have asymptotically valid *unconditional* coverage for the LATE, and expected length lower than usual CIs based on the TSLS estimator — reflecting the gains in terms of inference. Notice that the asymptotic study of CIs lengths requires to rescale CIs by $\sqrt{n}$ to allow for a meaningful comparison.[21]

COROLLARY 1. *Under assumptions 1 and 2, and for a common estimation sample size for both $\hat{\tau}_T$ and $\hat{\tau}^{TSLS}$, we have:*

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

---

[21]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 $\sqrt{n}$ rate).

2. $\lim_{n \to \infty} E\left[\sqrt{n} \cdot \text{length}[CI_\alpha(\hat{S})]\right] < \lim_{n \to \infty} \sqrt{n} \cdot \text{length}[CI_\alpha^{TSLS}]$

where expectations are taken with respect to the distribution of $\hat{S}$.

Proof. See appendix A.

$\square$

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 implicitly 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.[22] Indeed, the sample size used for estimation is divided by 2 (doubling the variance of the estimator all else equal), yet the compliance rate is doubled, dividing by 4 the variance. Yet in general, the researcher won't have an additional separate sample to implement the testing step. In this case, she will need to (randomly) split her sample in two sub-samples to implement our methodology, reducing the size of the estimation sample in comparison to the usual TSLS estimation. Suppose she implements a 20%-80% split to create a

---

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

test and estimation sample.[23] Then instead of dividing by two the size of the estimation sample (post-selection), she ends up reducing it by a factor of $\frac{4}{5} \cdot \frac{1}{2} = \frac{2}{5} < \frac{1}{2}$ — compared to the sample size used in TSLS estimation. Hence the reduction in variance goes from a factor of $\frac{1}{2}$ to a factor of $\frac{5}{2} \cdot \frac{1}{4} = \frac{5}{8} > \frac{1}{2}$. More generally, if the gains in variance derived from the increased compliance rate in the selected population aren't large enough, they can be cancelled by the losses due to the sample split — to the point that the overall procedure might lead to an *inflated* variance in the worst cases.

Cross-fitting The example above makes it clear that sample split is not innocuous. Yet this is a key step of our approach as it allows to make the testing-selecting and estimation steps independent. As already discussed in Abadie et al. (2022), this is key to avoid the emergence of a dependence between first-stage and second stage errors. Intuitively, if we were to test-select and estimate in the same sample, the selected groups would sometimes include groups with no compliers, for which by chance a disproportionate amount of always-takers ended up assigned to the treatment arm in our sample, leading the t-test to conclude to a significant first-stage.

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

---

[23]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.

used for selection. A sufficient condition for such gains in variance to be restored is that the two cross-fit estimators are independent one from another. This is what the following lemma establishes.

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

The above results are encouraging as they suggest that there are indeed potential 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.[24] In particular, our framework so far ruled out the possibility to wrongly exclude groups with some compliers — as by consistency of the t-test against any (well-separated from 0) alternative, the probability to exclude such groups from the selected sample was asymptotically zero. This is not a satisfactory approximation of what would happen in finite samples — in which groups with small shares of compliers might be wrongly excluded by the selection procedure — and 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). *There are groups with a local-to-zero share of compliers. Formally:*

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

---

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

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

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

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

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 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, 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: $\forall g \in \mathcal{G}_S, \ S_g = 1$. 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.

PROPOSITION 2. *Under assumptions 1 and 3, we have:*

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

2. $B(S) \propto |LATE^{W(S)} - LATE|$, where $LATE^{W(S)}$ denotes the average treatment effect among compliers within groups with weak first-stages that are wrongly dropped by the selection procedure S

3. $B(S) \neq 0$ if $\exists j$ s.t. $\{S_j = 0 \cap j \in \mathcal{G}_W\}$ and $LATE^{W(S)} \neq LATE$.

Proof. See appendix A. ☐

Without further assumptions on treatment effect heterogeneity, the above proposition suggests that our proposed estimator will systematically be biased in the presence of groups with weak first-stages — as the probability of wrongly excluding those groups does not vanishes to 0 asymptotically (see lemma 4) and proposition 2.3 shows that in the presence of such exclusion errors, the bias of our procedure is non-zero.

COROLLARY 2. *Under assumptions 1 and 3 and homogeneous treatment effect, we have that $\hat{\tau}(S)$ is unbiased and asymptotically normal:*

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

On the other hand, assuming homogeneous treatment effect is not realistic either, and rather in opposition to the spirit of the LATE literature. Inspired by Coussens and Spiess (2021), we follow a middle-ground by assuming below that treatment effect heterogeneity is of the same order of magnitude as sampling variation, i.e., decreasing at a $1/\sqrt{n}$ rate.

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:*

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

Coussens and Spiess (2021) justify this type of assumption by highlighting the usual difficulties in detecting treatment effect heterogeneity in empirical research, given the usual sample sizes. Let us note that contrary to Coussens and Spiess (2021), we do not assume that treatment effects in general are 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. The next theorem studies the asymptotic distribution of our estimator in such a framework.

THEOREM 1. *Under assumptions 1, 3 and 4, we have:*

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

2. $\sqrt{n_E} \cdot \dfrac{\hat{\tau}_T - LATE}{\sqrt{\hat{V}(\tau(\hat{S}_T))}} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 1)$

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.[25] 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

---

[25]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.

is less important than in Coussens and Spiess (2021), and the restriction on heterogeneity made in assumption 4 is sufficient to rule out any first-order bias. We view such a discrepancy in the behavior of our estimator compared to the one of Coussens and Spiess (2021) as revealing two points:

1. the heterogeneity restriction made in assumption 4 is far from being equivalent to homogeneous treatment effects, as estimators such as the one of Coussens and Spiess (2021) that would converge to the LATE in the homogeneous case exhibit a first-order bias under this assumption ;

2. our estimator offers gains in variance while remaining more tightly related to the LATE parameter than the one studied in Coussens and Spiess (2021). Hence we offer another alternative in the bias-variance trade-off, from no asymptotic bias (yet larger variance) when using TSLS to potentially larger gains in variance when using Coussens and Spiess (2021) (at the cost of a larger asymptotic bias, even under restrictions on treatment effect heterogeneity).

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

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

where we assumed $\mathrm{Var}[Y(0)] = \mathrm{Var}[Y(1)] = \sigma^2$,[26] and q_x is the xth-quantile of a $\mathcal{N}(0,1)$. Hence in studies designed based on power analyses, we have by design: $e^* = O(n^{-1/2})$. It can still be that the true effect (and treatment effect heterogeneity) is way larger than e^*, in which case our study

---

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

will systematically detect the effect of the policy (and its heterogeneity). This would be the case in general in sciences that are over-powered... yet social sciences (and economics in particular) have rather been documented to be *under*-powered in meta-analyses — e.g., in Ioannidis et al. (2017). Experimenters in social sciences certainly do not detect 100% of the time significant effects (and even less often treatment effect *heterogeneity*). Hence it might seem reasonable to assume that most of the time, the true effects (and true heterogeneity) is of the same order of magnitude as the MDE of the study designed to detect them. In such a case, assumption 4 would be fulfilled.

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:

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

where $G = 2$ denotes the population not selected, $G = 1$ the selected population, and $LATE^1$ and $LATE^2$ the LATEs within those two populations — i.e., $LATE^1$ denoted the estimand targeted by our procedure. Of course, $G = 1$ and $G = 2$ depend on the realization of the sample. Let us consider a sensitivity analysis that would condition on the sample realization, so that $G = 1$ and $G = 2$ are considered as deterministic.[27] The quantity $\mathrm{P}[G = 2 | D(1) > D(0)]$ can be estimated by

---

[27]In other words, $LATE^1$ and $LATE^2$ become estimands that are sample-dependent. This is not an issue as ulti-

2SLS as suggested in Abadie (2003).[28] 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 $(1-\alpha)$-confidence interval around this parameter — e.g., using inversion of an Anderson-Rubin statistic. Suppose we construct 99%-CI around $\mathrm{P}[G = 2|D(1) > D(0)]$, and take the upper bound of this quantity, denoted $\widehat{UB}^P$. The bias term B is increasing in $\mathrm{P}[G = 2|D(1) > D(0)]$, hence the *worst-case* bias can be obtained by replacing $\mathrm{P}[G = 2|D(1) > D(0)]$ with $\hat{UB}^P$. We are left with the unknown $\left(LATE^2 - LATE^1\right) \equiv M$, that is going to be our sensitivity parameter. For a given value of M, the worst-case bias of our proposed estimator for the LATE is given by $M \cdot \widehat{UB}^P$. Suppose we widen our 96%-CI around $LATE^1$ — the effect among compliers in the population selected by our procedure — by $\pm M \cdot \widehat{UB}^P$. Such CI is (asymptotically) valid with a 95% coverage for the overall LATE parameter.[29] 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 $s(X) \equiv \mathrm{E}[D(1) - D(0)|X]$, denoted $\hat{s}(X)$

2. No assumption on the rate of convergence of $\hat{s}(X)$. The hope is merely that $\hat{s}(X)$ contains some signal for the true $s(X)$.

---

mately, this sensitivity analysis will still be related to an estimand that is sample-independent, namely the LATE in the whole population.

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

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

3. Define $\bar{G}$ (a fixed number) groups based on quantiles of $\hat{s}(X)$, and use Chernozhukov et al. (2021) results to make inference on:

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

That way, we are back to a situation in which the covariates are reduced to a partition of the population: $\{\mathbb{I}_{\hat{s}(x) \in [q_{g-1}, q_g]}\}_{g \in \{1, \dots, G\}}$. Unfortunately, the procedure proposed in Chernozhukov et al. (2021) cannot be directly applied to our setting since it is based on repeated data-splits, with $\hat{s}(X)$ being estimated repeatedly, such that inference on so-called GATEs — grouped average treatment effects, of the form $\mathrm{E}[s(X)|\hat{s}(X) \in [q_{g-1}, q_g]]$ — can be made without offering a clear way to associate a given observation to a given group — since such a mapping will change from one data-split to another. A simple fix to this issue is to commit to a single data-split. This is the choice we make in our application in section 6. Chernozhukov et al. (2021) rightfully advocate against such approach — and defend instead the variational inference approach they develop — as (i) this could lead to 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 expose researchers to the risk of drawing a "bad" split. That said, 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 $\mathrm{Pr}[\text{sh. of compliers} = 0|g \in \mathcal{G}_S]$ that goes to 0, hence a weight that goes to 1 as is already the case in our proposed estimation strategy.

5 SIMULATIONS

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

DGP parameters   To simulate a flexible DGP, we use the threshold crossing model representation (Vytlacil, 2002).[30] Let $\delta_i, \varepsilon_i \sim \mathcal{N}\left(\vec{0}, \Sigma\right)$, with

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

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

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

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

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

We also define a covariate X as:

$$X_i = \delta_i + \eta_i$$

---

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

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

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

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

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

$$\sigma_{Y(1)-Y(0)}^2 = \beta^2 \sigma_{\pi_G}^2 \left(\alpha^2 + (1-\alpha)^2\right)$$
$$\mathrm{cov}(\pi_G, Y_i(1) - Y_i(0)) = \beta \cdot \alpha \cdot \sigma_{\pi_G}^2$$
$$\mathrm{cor}(\pi_G, Y_i(1) - Y_i(0)) = \frac{1}{\sqrt{1 + \left(1 - \frac{1}{\alpha}\right)^2}}$$

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,[31] 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

---

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

4. $\rho_{\delta\varepsilon}$: correlation between latency to treat and baseline untreated potential. Controls selection into treatment and hence the necessity for instrumentation.

5. σ_η^2: Controls how good of a predictor the groups are for compliance

6. α, β: Control the dependence between treatment effect and compliance as well as the overall treatment effect heterogeneity

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

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

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

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

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

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

We run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 1. We vary treatment effect heterogeneity, and quantify the latter on the x-axis by scaling the standard deviation of $Y(1) - Y(0)$ by the Minimum Detectable Effect (MDE):
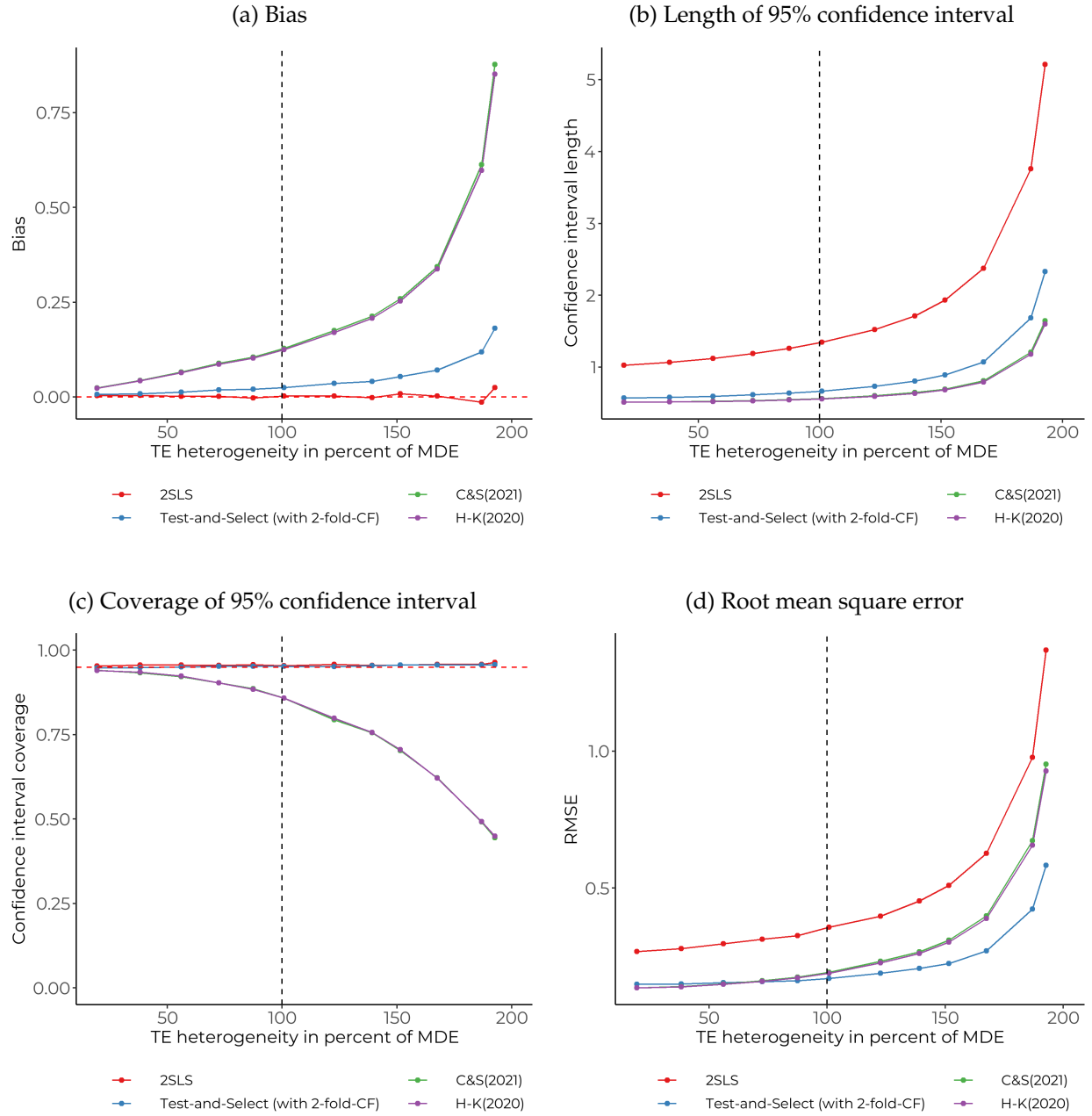
$$x = \frac{\sqrt{V(Y(1) - Y(0))}}{\sqrt{\frac{V(Y(1)) + V(Y(0))}{0.5 \cdot n}} \cdot \frac{q_{0.975} + q_{0.8}}{\pi}}$$

where q_x represents the quantile function of a normal. This re-scaling allows a meaningful quantification of treatment effect heterogeneity, by relating it to a quantity (the MDE) that (i) is a well-known object to most empiricists and (ii) varies with the sample size at a $1/\sqrt{n}$ rate. Recall that 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

(a) Bias

(b) Length of 95% confidence interval

(c) Coverage of 95% confidence interval

(d) Root mean square error

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

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

$$\text{DGP2} \equiv \left(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\delta\varepsilon} = 0.3, \sigma_\eta = 0.5, \alpha = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\} \right)$$

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

$$\pi_G = (\pi_1 = \pi_{10} \approx 0.001, \pi_2 = \pi_9 \approx 0.08, \pi_3 = \pi_8 \approx 0.24, \pi_4 = \pi_7 \approx 0.40, \pi_5 = \pi_6 \approx 0.5)$$
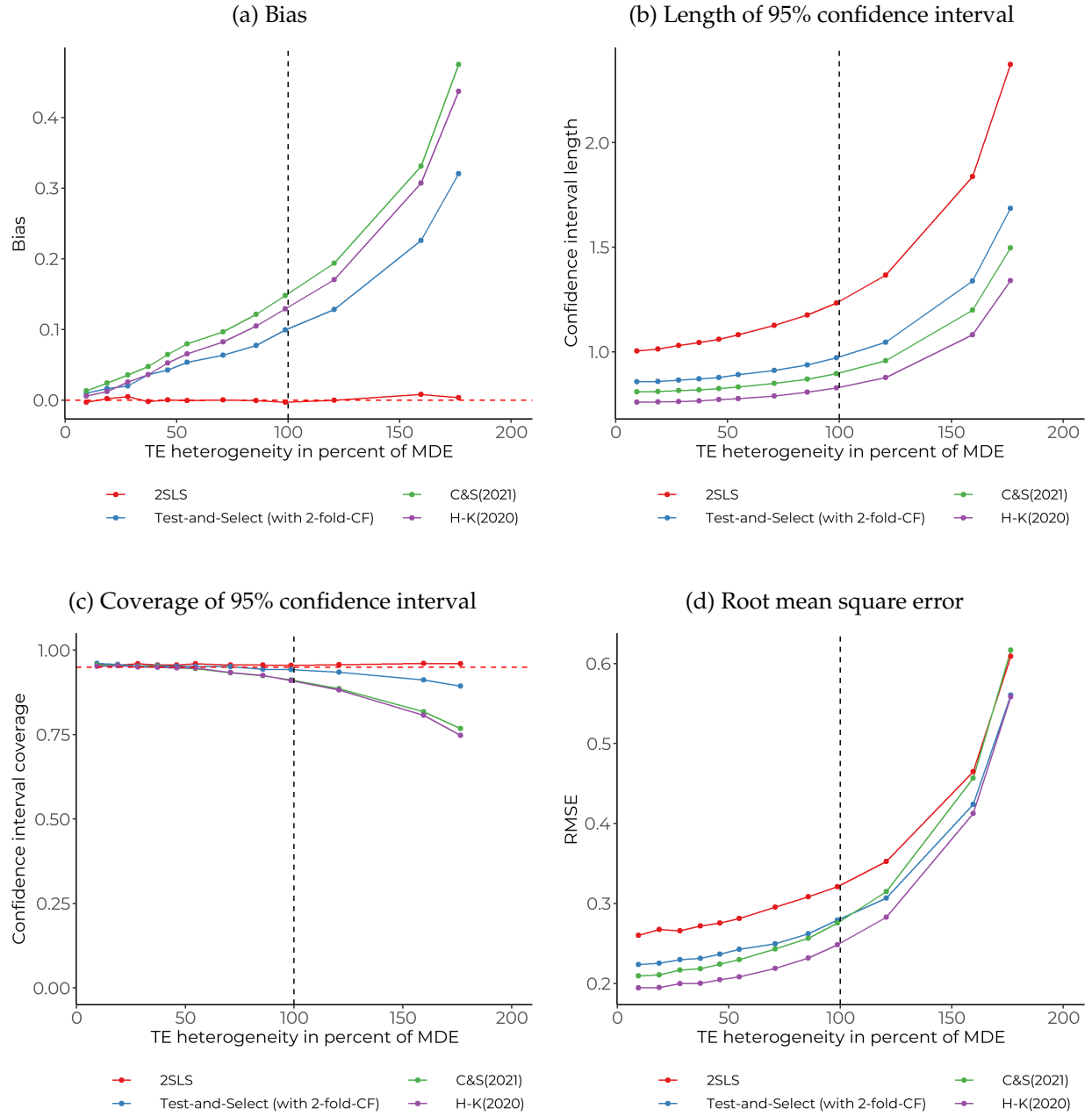
for an overall compliance of 25%, as in DGP1. In the spirit of varying parameters as little as possible across DGPs, we keep the same $\alpha = 0.5$ as in DGP1, which again means that there is a significant correlation between compliance and treatment effects across groups.

We run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 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.[32]

---

[32]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

(a) Bias

(b) Length of 95% confidence interval

(c) Coverage of 95% confidence interval

(d) Root mean square error

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 APPLICATION

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.[33] 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 our analysis. This data-splitting and cross-fitting allows us to consider this covariate as essentially

---

[33]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.

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 2 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 3 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 relatively larger deviation of their point estimates from the 2SLS one (compared to our method-

Table 2: 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.

ology). 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 2 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 3 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 3: 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.

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

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

Joshua D. Angrist and William N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *The American Economic Review*, 88(3):450–477, 1998. ISSN 00028282. URL http://www.jstor.org/stable/116844.

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.

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.

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.

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

A PROOFS OF MAIN RESULTS

Proof. **Lemma 1.**

First of all, notice that group $G = 1$ is selected with probability tending to 1 as n goes to infinity (by consistency of the t-test against alternatives well separated from 0). With probability tending to $(1 - \alpha)$ — where α is the level of the t-test used for selection — group $G = 0$ is not selected.

When group 1 alone is selected — with probability tending to $(1 - \alpha)$, then the first-stage parameter to be estimated is π^1. Since the event determining whether group 1 is selected along or not does not depend on the observations in this group, the first-stage estimator among observations with $G = 1$ will be $\sqrt{n}$-consistent for π^1 in such cases where only group G=1 is selected.

When both groups are selected, the first-stage parameter to be estimated is $FS \equiv \pi^0 \cdot \mathrm{P}[G = 0] + \pi^1 \cdot \mathrm{P}[G = 1]$. The difference in means estimator for this parameter can be decomposed as:

$$\widehat{FS} = \hat{\pi}^0 \cdot \widehat{\mathrm{P}[G = 0]} + \hat{\pi}^1 \cdot \widehat{\mathrm{P}[G = 1]}$$

$\widehat{\mathrm{P}[G = 0]}$ and $\widehat{\mathrm{P}[G = 1]}$ are the empirical analogs of $\mathrm{P}[G = 0]$ and $\mathrm{P}[G = 1]$, and consistent by the law of large numbers. By the same argument as above, $\hat{\pi}^1$ is a consistent estimator of π^1. On the contrary, $\hat{\pi}^0$ is not independent of the event {Group G=0 is selected}. Indeed, we have:

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

where $\hat{\sigma}^{\hat{\pi}^0}$ is a consistent estimator of the asymptotic variance of $\hat{\pi}^0$. Therefore we get:

$$\sqrt{n^0} \cdot \left(\hat{\pi}^0 - \underbrace{\pi^0}_{=0} \right) \geq \sqrt{n^0} \cdot \left(\frac{\hat{\sigma}^{\hat{\pi}^0}}{\sqrt{n^0}} \cdot q_{1-\alpha} - 0 \right)$$

$$= \hat{\sigma}^{\hat{\pi}^0} \cdot q_{1-\alpha}$$

$$\xrightarrow[n\to\infty]{} \sigma^{\hat{\pi}^0} \cdot q_{1-\alpha} \neq 0$$

Therefore:

$$\sqrt{n} \cdot (\widehat{FS} - FS) = \sqrt{n} \cdot \left(P[G=0] \cdot (\hat{\pi}^0 - \pi^0 + P[G=1] \cdot (\hat{\pi}^1 - \pi^1) + o_p(1) \right)$$

$$\xrightarrow[n \to \infty]{} \sqrt{P[G=0]} \cdot \sigma^{\hat{\pi}^0} \cdot q_{1-\alpha} \neq 0$$

Hence with probability $(1 - \alpha)$, the first-stage parameter is estimated with first-order bias. □

Proof. **Proposition 1.**

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

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

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

We do not consider groups with weak first-stages at this point, as proposition 1 focuses on standard asymptotics in order to illustrate the potential for gains in precision from selection.

Notice that Propositions 1.1 and 1.2 are developed under a conditioning on the value of the selection vector S. This is key to our reasoning, as this conditioning allows us to study separately the randomness of the estimation sample, and the one coming from the selection step.

Consider a given (fixed, deterministic) selection process $S \in \mathcal{S}_{\text{strong}}$. We know that asymptotically, it cannot be that a group with a strong first-stage is not selected. Hence there are only two main cases we need to consider:

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

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

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

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

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

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

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

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

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

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

$$\text{LATE} = \ \text{E}[Y(1) - Y(0)|D(1) > D(0), S_G = 1] \cdot \overbrace{\text{P}[S_G = 1|D(1) > D(0)]}^{=1}$$

$$+ \text{E}[Y(1) - Y(0)|D(1) > D(0), S_G = 0] \cdot \underbrace{\text{P}[S_G = 0|D(1) > D(0)]}_{=0}$$

$$= \text{E}[Y(1) - Y(0)|D(1) > D(0), S_G = 1]$$

$$= \frac{E[Y|Z=1, S_G=1] - E[Y|Z=0, S_G=1]}{E[D|Z=1, S_G=1] - E[D|Z=0, S_G=1]} \quad \text{(by standard identification result for LATE)}$$

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

$$a_i = \frac{Z_i S_{G_i}(Y_i - E[Y|Z=1, S_G=1])}{E[ZS]}$$

$$b_i = \frac{(1-Z_i) S_{G_i}(Y_i - E[Y|Z=0, S_G=1])}{E[(1-Z)S_G]}$$

$$c_i = \frac{Z_i S_{G_i}(D_i - E[D|Z=1, S_G=1])}{E[ZS_G]}$$

$$d_i = \frac{(1-Z_i) S_{G_i}(Y_i - E[D|Z=0, S_G=1])}{E[(1-Z)S_G]}$$

Therefore we get:

$$\psi_{\hat{\tau}(S),i} = \frac{(a_i - b_i) - LATE \cdot (c_i - d_i)}{C_i - D_i}$$

$$= \frac{1}{p_{C,S_G=1}} \left(\frac{Z_i S_{G_i}(Y_i - E[Y|Z=1, S_G=1])}{E[ZS_G]} - \frac{(1 - Z_i)S_{G_i}(Y_i - E[Y|Z=0, S_G=1])}{E[(1-Z)S_G]} \right.$$

$$\left. -LATE \cdot \left(\frac{Z_i S_{G_i}(D_i - E[D|Z=1, S_G=1])}{E[ZS_G]} - \frac{(1-Z_i)S_{G_i}(D_i - E[D|Z=0, S_G=1])}{E[(1-Z)S_G]} \right) \right)$$

$$= \frac{1}{p_{C,S_G=1}} \left(\frac{1}{E[ZS_G]} Z_i S_{G_i} \cdot (\varepsilon_i - E[\varepsilon|Z=1, S_G=1]) - \frac{1}{E[(1-Z)S_G]}(1-Z_i)S_{G_i} \cdot (\varepsilon_i - E[\varepsilon|Z=0, S_G=1]) \right)$$

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

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

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

$$V(\psi_{\hat{\tau}(S),i}) = E[\psi_{\hat{\tau}(S),i}^2]$$

$$= \frac{1}{p_{C,S_G=1}^2} \left(\frac{1}{E[ZS_G]} E[(\varepsilon - E[\varepsilon|Z=1, S_G=1])^2|Z=1, S_G=1] \right.$$

$$\left. + \frac{1}{E[(1-Z)S_G]} E[(\varepsilon - E[\varepsilon|Z=0, S_G=1])^2|Z=0, S_G=1] \right)$$

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

$$E[ZS_G] = p \cdot p_{S_G}$$

$$E[(1-Z)S_G] = (1-p) \cdot p_{S_G}$$

$$\pi = p_{C,S_G=1} \cdot p_{S_G} + p_{C,S_G=0} \cdot (1 - p_{S_G}) = p_{C,S_G=1} \cdot p_{S_G}$$

$$\Rightarrow V(\psi_{\hat{\tau}(S),i}) = \frac{p_{S_G}}{\pi^2} \left(\frac{1}{p} E[(\varepsilon - E[\varepsilon|Z=1, S_G=1])^2|Z=1, S_G=1] \right.$$

$$\left. + \frac{1}{1-p} E[(\varepsilon - E[\varepsilon|Z=0, S_G=1])^2|Z=0, S_G=1] \right)$$

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

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

$$V^{TSLS} = \frac{1}{\pi^2} \left(\frac{1}{p} V[\varepsilon|Z=1] + \frac{1}{1-p} V[\varepsilon|Z=0] \right)$$

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

Therefore, we only need to prove that:

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

This is proven below:

$$V(\varepsilon|Z=z) = E\left[V(\varepsilon|Z=z, S_G)|Z=z\right] + V(E\left[\varepsilon|Z=z, S_G\right]|Z=z)$$

$$\geq E\left[V(\varepsilon|Z=z, S_G)|Z=z\right]$$

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

where the first equality follows from the law of total variance, and first and second inequalities follow from the fact that variances are always positive or null (in degenerate cases).

Therefore, V^{TSLS} has been shown to be a linear combination (with positive coefficients) of terms greater or equal than the ones appearing in $V(\hat{\tau}(S))$, proving the proposition 1.2.

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

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

where $n_{(T)}^g$ is the number of observations in group g in sample $\mathcal{I}_T$, $\hat{\pi}_{(T)}^g$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance

of $\hat{\pi}^g_{(T)}$. Notice that $\hat{\pi}^g_{(T)}$ is asymptotically linear, as following lemma 6 we have:

$$\sqrt{n^g_{(T)}} \cdot [\hat{\pi}^g - \pi g]$$

$$= \sqrt{n^g_{(T)}} \cdot \left[\frac{\sum_i Z_i D_i}{\sum_i Z_i} - \frac{\sum_i (1 - Z_i) D_i}{\sum_i (1 - Z_i)} - (\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]) \right]$$

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

$$= \frac{1}{\sqrt{n^g_{(T)}}} \cdot \sum_{i=1}^{n^g_{(T)}} \tilde{\psi}^g_i$$

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

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

where the expression of the influence function is given by:

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

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

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

where $\hat{V}(\tau(\hat{S}_T))$ is a consistent estimator of the variance of $\tau(\hat{S}_T)$, the estimator conditional on the value of the selection vector. Now, turning to the study of the characteristic function of $\hat{T}_E$ conditional on $\hat{S}_T$, we have:

$$\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T] = \sum_{S \in \{0,1\}^{|\mathcal{G}|}} \mathbb{1}\{\hat{S} = S\} \cdot \mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T = S]$$

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

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

Notice that by Jensen inequality: $|\mathrm{E}[e^{it\hat{T}_E}|\hat{S}_T]| \leq \mathrm{E}[|e^{it\hat{T}_E}||\hat{S}_T] = 1$, hence by the dominated convergence theorem we get:

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

$$= e^{-t^2/2} \quad \text{(characteristic function of a } \mathcal{N}(0,1) \text{)}$$

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

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

By Jensen inequality again, we have:

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

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

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

$\square$

Proof. **Corollary 1**.

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

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

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

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

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

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

which is the first statement of corollary 1.

Regarding the second statement, proposition 1.2 states that

$$\forall S \in \mathcal{S}_{\text{strong}}, \; V(\hat{\tau}(S)) \leq V^{\text{TSLS}}$$

Once $\sqrt{n}$-rescaled, the length of $CI_\alpha(S)$ and CI_α^{TSLS} is entirely governed by (and increasing in) $\hat{V}(\hat{\tau}(S))$ and $\hat{V}(\hat{\tau}^{TSLS})$ respectively, which are consistent estimators of $V(\hat{\tau}(S))$ and V^{TSLS}. Indeed:

$$\forall S \in \mathcal{S}_{strong}, \quad \sqrt{n} \cdot \text{length}[CI_\alpha(S)] = 2 \cdot \sqrt{\hat{V}(\hat{\tau}(S))} \cdot q_{1-\frac{\alpha}{2}}$$
$$\sqrt{n} \cdot \text{length}[CI_\alpha^{TSLS}] = 2 \cdot \sqrt{\hat{V}(\hat{\tau}^{TSLS})} \cdot q_{1-\frac{\alpha}{2}}$$

Hence proposition 1.2 implies:

$$\forall S \in \mathcal{S}_{\text{strong}}, \; \lim_{n\to\infty} \text{length}[\sqrt{n} \cdot CI_\alpha(S)] \leq \lim_{n\to\infty} \text{length}[\sqrt{n} \cdot CI_\alpha^{TSLS}]$$

Thus we have:

$$\lim_{n\to\infty} \mathrm{E}[\text{length}[\sqrt{n}\cdot CI_\alpha(\hat{S})]] = \lim_{n\to\infty} \big[\mathrm{E}[\text{length}[\sqrt{n}\cdot CI_\alpha(S)]|S\in\mathcal{S}_{strong}]\cdot\mathrm{P}[S\in\mathcal{S}_{strong}]$$

$$+\,\mathrm{E}[\text{length}[\sqrt{n}\cdot CI_\alpha(S)]|S\notin\mathcal{S}_{strong}]\cdot\mathrm{P}[S\notin\mathcal{S}_{strong}]\big]$$

$$=\lim_{n\to\infty}\mathrm{E}[\text{length}[\sqrt{n}\cdot CI_\alpha(S)]|S\in\mathcal{S}_{strong}]$$

$$\leq\lim_{n\to\infty}\text{length}[\sqrt{n}\cdot CI_\alpha^{TSLS}]$$

where the last equality comes from the consistency of t-test that implies $\mathrm{P}[S\in\mathcal{S}_{strong}]\xrightarrow[n\to\infty]{}1$, and the inequality (which is the second statement of the corollary) comes from the inequality obtained in the last but one display. $\qquad\square$

Proof. **Lemma 3**.

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

First of all, if we denote by $\hat{\tau}_1$ the estimator constructed using the fold $\mathcal{I}_2$ as the test sample and $\mathcal{I}_1$ as the estimation sample, recall that we can decompose it as follows:

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

where the expression of the influence function is given by:

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

$$\left.-\frac{1}{E[(1-Z)\hat{S}_{G,(2)}]}(1-Z_i)\hat{S}_{G_i,2}\cdot(\varepsilon_i - E[\varepsilon|Z=0,\hat{S}_{G,(2)}=1])\right)$$

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

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

where $n_{(2)}^g$ is the number of observations in group g in sample $\mathcal{I}_1$, $\hat{\pi}_{(2)}^g$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance

52

of $\hat{\pi}^g_{(2)}$. Second, recall (from the proof of corollary 1 above) that:

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

$$
= \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i
$$

The above formulas make it clear that the potential source of dependence between $\hat{\tau}_1$ and $\hat{\tau}_2$ lies in $\hat{S}_{g,(2)}$, that appears in the influence function of $\hat{\tau}_1$ and is computed based on observations from fold $\mathcal{I}_2$, also used in $\hat{\tau}_2$. We will now study the (asymptotic) dependence of $\hat{S}_{g,(2)}$ on $\tilde{\psi}^g_n$, the nth individual influence function entering in $\hat{\pi}^g_{(2)}$. For groups g such that $\pi^g > 0$ (strong first-stage), we have that $\mathrm{P}[\hat{S}_{g,(2)} = 1] \xrightarrow[n \to \infty]{} 1$ and $\hat{S}_{g,(2)}$ becomes essentially deterministic, hence asymptotically there aren't any dependence issues for such groups. We will therefore focus on groups g such that $\pi^g = 0$. For any such group g, and for a given number of observations $n^g_{(2)}$ in this group (in fold $\mathcal{I}_2$), we have:

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

$$
= \mathbb{1} \left\{ \left| \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i \right| > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n^g_{(2)}}} \cdot q_{1-\frac{\alpha}{2}} \right\}
$$

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

where we defined: $F^{g,n^g_{(2)}} \equiv \frac{1}{\sqrt{n^g_{(2)}}} \cdot \sum_{i=1}^{n^g_{(2)}} \tilde{\psi}^g_i$. Hence we can study the probability that any addi-

tional observation modifies the value of $\hat{S}_{g,(2)}^{(n_{(2)}^g)}$ as follows:

$$P\left[\hat{S}_{g,(2)}^{(n_{(2)}^g-1)} = 0, \ \hat{S}_{g,(2)}^{(n_{(2)}^g)} = 1\right]$$

$$= P\left[\left|F^{g,n_{(2)}^g-1}\right| \leq \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n_{(2)}^g - 1}} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right), \quad \left|F^{g,n_{(2)}^g}\right| > \frac{\hat{\sigma}^{\pi^g}}{\sqrt{n_{(2)}^g}} \cdot q_{1-\frac{\alpha}{2}}\right]$$

$$= P\left[\left|\left(n_{(2)}^g - 1\right) \cdot F^{g,n_{(2)}^g-1}\right| \leq \sqrt{n_{(2)}^g - 1} \cdot \hat{\sigma}^{\pi^g} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right), \quad n_{(2)}^g \cdot \left|F^{g,n_{(2)}^g}\right| > \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}}\right]$$

Notice that:

$$n_{(2)}^g \cdot \left|F^{g,n_{(2)}^g}\right| = \left|\tilde{\psi}_{n_{(2)}^g}^g + \frac{1}{\sqrt{n_{(2)}^g}} \cdot \sum_{i=1}^{n_{(2)}^g-1} \tilde{\psi}_i^g\right|$$

$$= \left|\tilde{\psi}_{n_{(2)}^g}^g + \left(n_{(2)}^g - 1\right) \cdot F^{g,n_{(2)}^g-1}\right|$$

$$\leq \left|\tilde{\psi}_{n_{(2)}^g}^g\right| + \left(n_{(2)}^g - 1\right) \cdot \left|F^{g,n_{(2)}^g-1}\right| \qquad \text{(by the triangle inequality)}$$

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

$$P\left[\hat{S}_{g,(2)}^{(n_{(2)}^g-1)} = 0, \ \hat{S}_{g,(2)}^{(n_{(2)}^g)} = 1\right]$$

$$= P\left[\left|\left(n_{(2)}^g - 1\right) \cdot F^{g,n_{(2)}^g-1}\right| \leq \sqrt{n_{(2)}^g - 1} \cdot \hat{\sigma}^{\pi^g} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right), \quad n_{(2)}^g \cdot \left|F^{g,n_{(2)}^g}\right| > \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}}\right]$$

$$\leq P\left[\left|\tilde{\psi}_{n_{(2)}^g}^g\right| > \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}} - \left(n_{(2)}^g - 1\right) \cdot \left|F^{g,n_{(2)}^g-1}\right|, \quad \left(n_{(2)}^g - 1\right) \cdot \left|F^{g,n_{(2)}^g-1}\right| \leq \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right)\right]$$

$$\leq P\left[\left|\tilde{\psi}_{n_{(2)}^g}^g\right| > \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}} - \left(n_{(2)}^g - 1\right) \cdot \left|F^{g,n_{(2)}^g-1}\right|, \right.$$

$$\left. \left(n_{(2)}^g - 1\right) \cdot \left|F^{g,n_{(2)}^g-1}\right| \leq \sqrt{n_{(2)}^g - 1} \cdot \hat{\sigma}^{\pi^g} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right)\right]$$

$$\leq P\left[\left|\tilde{\psi}_{n_{(2)}^g}^g\right| > \sqrt{n_{(2)}^g} \cdot \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}} - \sqrt{n_{(2)}^g - 1} \cdot \hat{\sigma}^{\pi^g} \cdot \left(q_{1-\frac{\alpha}{2}} - \epsilon\right)\right]$$

$$= P\left[\left|\tilde{\psi}_{n_{(2)}^g}^g\right| > \hat{\sigma}^{\pi^g} \cdot q_{1-\frac{\alpha}{2}} \cdot \left(\sqrt{n_{(2)}^g} - \sqrt{n_{(2)}^g - 1}\right) + \epsilon \cdot \sqrt{n_{(2)}^g - 1} \cdot \hat{\sigma}^{\pi^g}\right]$$

For $n_{(2)}^g$ large enough, we have:

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

Hence we get:

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

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

Proof. **Proposition 2**.

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

1. One including all groups with a strong or a weak first-stage, plus groups with zero first stages that are included in the selected sample defined by S. By construction, overall this group has a strong first-stage.

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

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

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

By proposition 1 and the argument above, we have:

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

Now, we turn to the case where S belongs to the set $\mathcal{S}_{\text{strong}}^1$. This includes all cases in which some

of the groups with a weak share of compliers get excluded from the restricted sample. We can always reframe such a situation by redefining two groups:

1. Group 1 including all selected groups as defined S. By construction, overall this group has a strong first-stage.

2. Group 2 including all excluded groups. By construction, since (by definition of $\mathcal{S}_{\text{strong}}^1$) 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.

$\square$

Proof. **Theorem 1**.

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

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

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

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

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

for triangular arrays (Lindeberg-Feller CLT) instead of the standard CLT — we get:

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

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

□

B PROOFS OF USEFUL LEMMAS

Proof. **Lemma 2**.

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

$$T^g_{\alpha, n_T} = \mathbb{1}\left\{ \left| \sqrt{n^g_T} \cdot \frac{\hat{\pi}^g}{\hat{\sigma}^g} \right| > q_{1-\frac{\alpha}{2}} \right\}$$

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

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

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

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

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

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

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

Proof. **Lemma 4**.

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

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

The quantiles of a $|\mathcal{N}(b, 1)|$ are increasing in b, and by assumption $H^g > 0$. Hence using the same

definition of the test statistic as in the proof of lemma 2, we get:

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

$\square$

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

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

and

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

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

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

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

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

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

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

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

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

where we went from the first to the second equality because (i) $(\hat{B} - B) = o_P(1)$ by the weak LLN, since it is an empirical mean of terms b_i with expectation 0, (ii) $\hat{B} = O_p(1)$ since it converges in probability to $B < \infty$, and (iii) since $O_P(1)^{-1} = O_P(1)$ and $o_P(1) \cdot O_P(1) = o_P(1)$, we have:

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

$\square$

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:* $\psi_i = \frac{Z_i(Y_i - E[Y|Z=1])}{E[Z]}.$

Proof.

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

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

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

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

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

$$= \frac{1}{\sqrt{n}}\frac{\sum_i Z_i(Y_i - E[Y|Z = 1])}{E[Z]} + o_P(1)$$

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 $\hat{B} = \sum_i Z_i$, $B = \mathrm{E}[Z]$, $b_i = Z_i - \mathrm{E}[Z]$. $\square$

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

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

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

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

Proof. The Wald estimator is a 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:

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

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

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

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

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

$$\Rightarrow \hat{\tau}^{Wald} = \frac{\hat{A}-\hat{B}}{\hat{C}-\hat{D}}$$

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

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

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

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

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

We then have:

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

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

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

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

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

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

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

$\square$

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:*

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

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

Proof. Let's consider a case with only three groups: a group with a strong first-stage ($\pi^1 > 0$), a

group with a weak first-stage ($\pi^2 = H^2/\sqrt{n}$), and a group with a zero first-stage ($\pi^3 = 0$).

Group 1 is always selected as asymptotically (as n_T goes to infinity), the selection procedure selects groups with a strong first-stage with probability 1.

Group 3 being selected or not does not affect the expectation of the limiting distribution of the ($\sqrt{n}-$scaled) resulting estimator, as shown in the proof of proposition 1.1. Hence we can ignore group 3 — or simply redefine group 1 or group 2 as including group 3 as well — without any changes in the result, and simply consider the two following cases:

1. Group 1 is selected, group 2 is selected

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

In the first case, the resulting estimator is the standard Wald estimator[34] computed on the whole estimation sample... hence it is $\sqrt{n}-$consistent (no asymptotic bias).

In the second case, the resulting estimator corresponds to the Wald estimator computed on group 1. Hence it is a $\sqrt{n}-$consistent estimator for the LATE conditional on being in group 1, which we define below:

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

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

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

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

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

At first, the quantities involved above might seem independent of n_E. The dependence of $LATE$ on n_E comes from the fact that the share of compliers in group 2 depends on n_E, as we have: $\pi^2 = H^2/\sqrt{n_E}$.

---

[34]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:

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

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

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

$$+ \ \mathrm{E}[Y(1) - Y(0)|D(1) > D(0), G = 2] \cdot \Pr[G = 2|D(1) > D(0)] \quad \text{(Law of iterated exp.)}$$

$$= LATE^1 \cdot \frac{\Pr[D(1) > D(0)|G = 1] \cdot \Pr[G = 1]}{\Pr[D(1) > D(0)]}$$

$$+ \ LATE^2 \cdot \frac{\Pr[D(1) > D(0)|G = 2] \cdot \Pr[G = 0]}{\Pr[D(1) > D(0)]} \quad \text{(Bayes' rule)}$$

$$= LATE^1 \cdot \frac{\pi^1 \cdot \Pr[G = 1]}{\pi} + LATE^2 \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi}$$

where the last line uses our standard notations:

$$\pi^g \equiv \mathrm{E}[D(1) - D(0)|G = g]$$

$$\pi \equiv \mathrm{E}[D(1) - D(0)] = \pi^1 \cdot \Pr[G = 1] + \pi^2 \cdot \Pr[G = 2]$$

Hence we get:

$$\sqrt{n_E} \cdot \left(LATE^1 - LATE\right) = \sqrt{n_E} \cdot \left(LATE^1 \cdot \left(1 - \frac{\pi^1 \cdot \Pr[G = 1]}{\pi}\right) - LATE^2 \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi}\right)$$

$$= \sqrt{n_E} \cdot \frac{\pi^2 \cdot \Pr[G = 2]}{\pi} \cdot \left(LATE^1 - LATE^2\right)$$

$$= \frac{H^2 \cdot \Pr[G = 2]}{\pi} \cdot \left(LATE^1 - LATE^2\right)$$

Therefore, we have in this case:

$$\sqrt{n_E} \cdot \left(\hat{\tau}(S^2) - LATE\right) = \sqrt{n_E} \cdot \left(\hat{\tau}(S^2) - LATE^1\right) + \sqrt{n_E} \cdot \left(LATE^1 - LATE\right)$$

$$= \sqrt{n_E} \cdot \left(\hat{\tau}(S^2) - LATE^1\right) + B(S^2)$$

$$\xrightarrow{d} \mathcal{N}(B(S^2), V^{S^2}) \quad \text{(Slutsky's lemma)}$$

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

$\square$

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

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

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

Their proof goes as follows:

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

where:

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

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

We simply need to study whether we still have:

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

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

Indeed, we have:

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

$$= \frac{\mathrm{E}[\alpha(X)w(X)\sqrt{n}\tau(X)]}{\mathrm{E}[\alpha(X)w(X)]} - \frac{\mathrm{E}[\alpha(X)\sqrt{n}\tau(X)]}{\mathrm{E}[\alpha(X)]}$$

$$= \frac{\mathrm{E}[\alpha(X)w(X)\mu(X)]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)\mu(X)]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}$$

$$- \frac{\mathrm{E}[\alpha(X)w(X)\sqrt{n}\mu]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)\sqrt{n}\mu]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}$$

$$= \frac{\frac{\mathrm{E}[\alpha(X)w(X)\mu(X)]}{\mathrm{E}[\alpha(X)]} - \frac{\mathrm{E}[\alpha(X)\mu(X)]}{\mathrm{E}[\alpha(X)]}\frac{\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]}}{\frac{\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]}} - \sqrt{n}\mu \underbrace{\frac{\mathrm{E}[\alpha(X)w(X)]\mathrm{E}[\alpha(X)] - \mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}{\mathrm{E}[\alpha(X)]\mathrm{E}[\alpha(X)w(X)]}}_{=0}$$

$$= B_w + 0$$

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:

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

We vary the number of groups as a way to exacerbate the pre-testing issue in simulations. We report in Table 4 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 4 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.[35]

Table 4: 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 4 for a varying number of

---

[35]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 4. 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
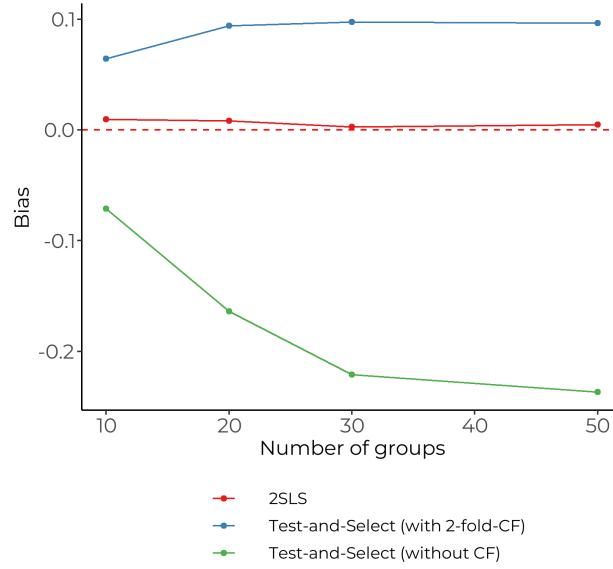
Notes: This figure shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP0, described in the text. Three different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select estimator with cross-fitting using 2 folds in blue, and a version of our 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.