IMPROVING LATE ESTIMATION IN EXPERIMENTS WITH IMPERFECT COMPLIANCE

YAGAN HAZARD
Collegio Carlo Alberto and ESOMAS, University of Turin

SIMON LÖWE

Estimation of causal effects in controlled or natural "experiments with imperfect compliance" usually relies on an Instrumental Variable strategy, which can yield imprecise and 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. We provide conditions under which it dominates the usual 2SLS estimator in terms of precision. Under an assumption on the degree of treatment effect heterogeneity, our estimator remains first-order unbiased with respect to the Local Average Treatment Effect estimand. We illustrate finite sample properties, robustness to treatment effect heterogeneity and precision gains using Monte Carlo simulations. Applying our methodology to the returns to schooling example, we document that a reduction in standard errors of 12% to 48% depending on specifications.

KEYWORDS: IV, LATE, imperfect compliance, precision, variance.

Yagan Hazard: yagan.hazard@carloalberto.org

Simon Löwe : loewe.sim@gmail.com

We thank Luc Behaghel and Xavier D'Haultfœuille for their invaluable support. Y. Hazard is deeply grateful to Toru Kitagawa, Jon Roth, Peter Hull, Soonwoo Kwon, Emily Oster and Jesse Shapiro for enlightening discussions at Brown University. We also thank Clément de Chaisemartin, Marc Gurgand, Philipp Ketz, Eric Maurin, Aleksey Tetenov and seminar participants at the PSE-CREST and Brown internal seminars for helpful comments and discussions. We gratefully acknowledge the support of the EUR grant ANR-17-EURE-0001. We provide a companion R package `late.rest` implementing our estimation strategy at https://github.com/simon-lowe/late.rest. All remaining errors are our own.

1. INTRODUCTION

Instrumental variables (IV) strategies are an integral part of the standard toolkit of applied economists and social scientists. This is due in part to their use for the estimation of causal effects in controlled or natural experiments with imperfect compliance. Such experiments are pervasive in applied research, since many interventions (such as education or training programs) cannot be imposed on a randomly selected group. Instead, in such cases, members of the treatment group are simply encouraged or given the opportunity to benefit from the intervention. However, 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.

However, a low *average* compliance rate can obscure highly heterogeneous compliance behaviors across subpopulations with different observable characteristics. This leaves room for researchers to improve the precision of their estimation by taking into account this heterogeneity. In this paper, we propose and study the properties of an intuitive way to take advantage of such heterogeneity. Our Test-and-Select estimator restricts IV estimation to subpopulations with significant non-zero compliance rates in sample. Excluding subgroups 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.

The present paper is structured as follows. We first underline the pitfalls of "naively" implementing such a selection rule based on estimated compliance rates, and then propose that data-splitting provides a simple fix to this issue. Next, we study the asymptotic properties of the Test-and-Select estimator under both standard and weak-IV-like asymptotic sequences. The former allows us to illustrate the potential gains in precision, while the latter aims at better approximating the finite sample properties of our proposed estimator. We also show that the Test-and-select procedure is robust to treatment effect heterogeneity. Indeed, we show that our proposed estimator remains first-order unbiased with respect to the Local Average Treatment Effect (LATE) under patterns of treatment effect hetero-

geneity that would generate a first-order bias in alternative estimation strategies proposed in the literature (Coussens and Spiess, 2021, Huntington-Klein, 2020, Abadie et al., 2022). Lastly, we study the finite sample properties of this estimator in Monte-Carlo simulations and in an application using changes in compulsory schooling as an instrument for education (Stephens and Yang, 2014). These sections illustrate (i) the potential gains in precision from implementing our methodology instead of the usual 2SLS estimator, and (ii) the improved robustness of our estimator to treatment effect heterogeneity compared to alternatives.

The burden placed by low compliance rates on the precision of the Two-Stage-Least-Squares (2SLS) estimator is well-known to most empiricists, and best illustrated by the variance formula of the 2SLS estimator in the simple case where the variance of the errors (denoted σ_ε^2) is homoscedastic.[1] Denoting by N the sample size, p the share of encouraged individuals, and π the share of compliers, we get:

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

A low compliance rate has a disproportionately large effect on the variance of the 2SLS estimator of the LATE, as the variance scales with the *square* of the compliance rate. Consider two experiments evaluating the same program, one with a 10% compliance rate ($\pi = 0.1$) and another with a perfect compliance ($\pi = 1$). Even though compliance rate in the first experiment is only 10 times lower than in the second experiment, a sample size a 100 times larger is needed to reach the same variance as in the perfect compliance experiment. Put differently, suppose it were possible in the first experiment to use some observables to identify the subpopulation of compliers. Focusing on this fraction (10%) of the population would divide the size of the estimation sample by 10, but it would still *decrease* the variance by a factor of 10, and thus significantly improve inference. In summary, even if a given experiment passes some weak identification tests successfully, which it could even with relatively low compliance rates, a low take-up rate can still significantly lower precision, possibly leading to uninformative inference.

As a concrete example, consider the quarter-of-birth instrument in Angrist and Krueger (1991). The instrument relies on the fact that because of compulsory schooling laws, chil-

[1] 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).

dren born at the beginning of the year are legally allowed to drop out earlier than those born at the end of the year, which leads the former to complete fewer years of schooling than the latter on average. However, preferences for education are likely to be highly heterogeneous along multiple dimensions such as parent's income and qualifications. It is for instance possible that children of richer parents never drop out in which case, their quarter-of-birth has no effect on their educational attainment. More generally, there might exist some subpopulations that do not react to the quarter-of-birth instrument, and would therefore not contribute to the identification of the LATE. Importantly, the existence of such non-compliant groups is not a threat to identification,[2] but their presence in the estimation sample does reduce the precision with which the LATE is estimated. It is intuitive to drop these groups without compliers from the estimation sample. This paper shows how to make this strategy operational and studies its properties.

Under "standard asymptotics",[3] which lead to perfect selection of groups without compliers, our estimator targets the same LATE parameter as the usual 2SLS/Wald estimator, while providing substantial precision gains. However, 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" for some groups,[4] because such asymptotics leave room for erroneous exclusions of groups with a non-zero share of compliers, even asymptotically. Without any restrictions on treatment effect heterogeneity, our proposed estimator has a first-order bias [5] for the LATE, as wrongly excluded groups could have an arbitrarily

[2] 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.

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

[4] 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.

[5] An estimator $\hat{\theta}$ of a parameter θ has a "first-order" or "asymptotic" bias when the limiting distribution of $\sqrt{n} \cdot \left(\hat{\theta} - \theta\right)$ is not centered on 0. This does not prevent $\hat{\theta}$ from being a consistent estimator of θ, but indicates that it does not converges towards θ at a $\sqrt{n}$-rate, which can invalidate inference based on such asymptotic approximation. Throughout the paper, we will use the term "$\sqrt{n}$-rate consistency" as synonymous to asymptotic unbiasedness, even if there terms are often times not used equivalently.

large treatment effect. We therefore provide conditions under which the estimand that our methodology targets is first-order equivalent to the LATE estimand. A sufficient condition for this to hold is for the the degree of treatment effect heterogeneity across groups to be of the same order of magnitude as the sampling variation. In other words, between-group heterogeneity is such that it would not systematically be detected in finite samples. We discuss in detail why this is a reasonable condition in practice. We also propose a data-splitting and cross-fitting strategy that provides valid inference despite the pre-test our estimation strategy relies on. We investigate the finite sample properties of our proposed procedure in Monte Carlo simulations. An R-package `late.rest` that implements our estimator (and allows for replication of our Monte-Carlo simulations) is available at https://github.com/simon-lowe/late.rest.[6]

Our paper contributes to the recent literature that has revisited IV estimation strategy in the presence of first-stage heterogeneity in order to achieve precision gains (Huntington-Klein, 2020, Coussens and Spiess, 2021, Abadie et al., 2022). Huntington-Klein (2020) and Coussens and Spiess (2021) consider interacted IV approaches, where the instrument is interacted with the covariates, which results in different weighting schemes. They show that this can yield significant improvements in precision and finite-sample bias at the cost of changing estimand to a convex weighted average of LATEs in the presence of treatment effect heterogeneity. Abadie et al. (2022) similarly consider an interacted IV approach but as in our proposed method additionally drop low-compliance groups. We differ from these papers in that we maintain the goal of estimating the standard LATE parameter instead of a weighted average of LATEs. Besides being more directly interpretable, the LATE parameter is also generally more directly policy-relevant parameter, as it corresponds to an existing subpopulation that can be targeted by policy-makers by using the exact same encouragement device (instrument) as in the experimental setting. Our pre-test for the presence of compliers in different subgroups of the population is also related to Bayesian approaches to inference on the LATE. Indeed, it has been previously noted that such approaches could yield more accurate estimates of causal parameters by specifically modelling the (heterogeneous) compliance behaviour along observable characteristics (Imbens and Rubin, 1997,

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

Rubin, 1998, Imbens and Rubin, 2015). While these methods fully model the compliance behavior, our procedure may be understood as a partial (and implicit) modelling of *non-compliance*.

Our paper also contributes to the pre-testing literature (Leeb and Pötscher, 2005), by introducing a correction method based on cross-fitting and proving some finite sample properties using a local-to-zero approach. We also contribute to the literature on semi-parametrically efficient estimation of the LATE parameter by pointing out that efficiency is lost when allowing for the empirically relevant case with zero or local-to-zero compliance in certain groups (Frölich, 2007, Hong and Nekipelov, 2010).

The remainder of the paper unfolds as follows. Section 2 presents the general framework and introduces the proposed estimator. Section 3 develops the theoretical results, in particular, the reduction in asymptotic variance achieved by our procedure compared to the usual 2SLS estimator. Section 4 studies the finite sample properties of our proposed estimation strategy, and compares it to alternatives. Section 5 presents an empirical application from a natural experiment using variation in compulsory schooling laws as an instrument for educational attainment. Section 6 then suggests some extensions. Lastly, section 7 concludes and presents some avenues for further research on this topic.

2. FRAMEWORK AND PROPOSED ESTIMATOR

We consider a data-generating process with a super-population $(Y(1), Y(0), D(1), D(0), Z, G)$, where $(Y(1), Y(0))$ are the potential outcomes when treated or not ($D = 1$ or 0), $(D(1), D(0))$ are the potential outcomes when encouraged or not ($Z = 1$ or 0), Z is the encouragement status, and G is a discrete pre-determined covariate (assumed binary in this section for illustrative purposes). We consider the simple yet standard case, especially in empirical work, where the treatment D and the instrument Z are binary. We then have:

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

We sample n independent and identically distributed observations $\{(Y_i, D_i, Z_i, G_i)\}_{1,\dots,n}$ from this super-population. We work under the standard identifying assumptions of the LATE (Angrist et al., 1996) stated below.

ASSUMPTION 1—LATE identifying assumptions:

1. Exclusion restriction: $Y(D,Z) = Y(D)$
2. Independence: $(Y(1), Y(0), D(1), D(0), G) \perp Z$[7]
3. First Stage: $\mathrm{E}[D = 1 \mid Z = 1] - \mathrm{E}[D = 1 \mid Z = 0] > 0$
4. Monotonicity: $D(1) \geq D(0)$

The only additional assumption compared to the standard framework is the independence of the covariates G and the instrument Z. This is trivially satisfied for any covariates determined prior to the draw of the instrument Z. Under this set of assumptions, Angrist et al. (1996) showed that the LATE, defined as the average treatment effect among compliers $\mathrm{E}[Y(1) - Y(0)|D(1) > D(0)]$, is identified. Compliers are individuals who respond to the instrument and therefore are defined by $D(1) > D(0)$. The usual estimator for the LATE is the Wald estimator (which coincides with the two-stage-least-squares (2SLS) estimator) in our case where D and Z are binary:

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

where E_n denotes the empirical mean operator.

Now 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 not affected by Z or D, as it is determined before the realization of Z and D. Take G as an indicator for right- ($G = 1$) or left-handedness($G = 0$). We allow for heterogeneous shares of compliers across the two groups, i.e., right-handed individuals might react more (or less) to the encouragement. For-

[7] In natural experiments, such assumption might only hold conditional on some observables. 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.

mally:

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

We do not impose that both π^0 and π^1 are strictly larger than 0, only that the average share of compliers in the population is well separated from 0 (assumption 1.3). In other words, if one of the two groups is fully unresponsive to the encouragement, the other needs to be responsive enough to allow for identification of the overall LATE. The existence of subpopulations with few compliers[8] (or no compliers at all) is what creates room for precision gains in the estimation of the overall LATE.

Consider the extreme case where left-handed share of compliers is $\pi^0 = 0$, while the right-handed share is $\pi^1 > 0$. The left-handed observations then do not bring any signal in the estimation of the overall LATE, as none of the compliers are left-handed:

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

They also do not prevent us from getting a consistent estimator of the LATE, as the difference in outcomes between encouraged and control left-handed individuals in the numerator of the usual LATE estimator cancels out on average, but they do bring additional noise to the estimation procedure, worsening the precision of the estimator:

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

where $\Delta_n^{\cdot|G=g} \equiv \mathrm{E}_n[\,\cdot\,|Z = 1, G = g] - \mathrm{E}_n[\,\cdot\,|Z = 0, G = g]$. This is easily seen when comparing the variance of a 2SLS estimator on the sample of right-handed individuals only ($\mathrm{V}^{TSLS,G=1}$) with the one on the full sample (V^{TSLS}), assuming homoscedasticity of

[8] We will formalize this vague terminology into the concept of a "local-to-zero" share of compliers, where the compliance rate vanishes at a $1/\sqrt{n}$ rate(Staiger and Stock, 1997).

the errors:

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

where σ_ε^2 denotes the variance of the errors,[9] N is the sample size and $p = E[Z]$ is the share of encouraged individuals. Excluding the group without compliers ($G = 0$) from the estimation decreases the variance by a factor $(1 - \Pr[G = 0])$. This is intuitive: the more we can get rid of groups without compliers, the larger the precision gains.

This suggests the following estimation procedure (Estimator 1), which we will call the "naive" Test-and-Select (naive TS) estimator.[10]

Algorithm 1 "Naive" Test-and-Select

1: For each group defined by G, do a one-sided t-test on the first-stage coefficient π^g, at a pre-specified level α, testing the null $\pi^g = 0$ against the alternative $\pi > 0$ (or $\pi < 0$).
2: Keep only the groups for which we rejected the previous test.
3: Compute the usual Wald 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 for which groups there are no compliers. Pre-testing can create challenges for inference (Leeb and Pötscher, 2005), and recent work underlined issues with the specific procedure of pre-testing on the first-stage in IV estimation (Abadie et al., 2022). The following lemma shows that pre-testing as suggested above and estimating in the same sample will lead to a first-order bias in the estimation of the LATE parameter.

[9] Here, ε is the structural error term in the "second stage" regression of the outcome on the treatment variable. Formally: $\varepsilon = Y - \mathrm{LATE} \cdot D$.

[10] Usually, in the context of RCTs, researchers will have a strong prior about the way their encouragement affects the treatment status, justifying $\pi > 0$ (or $\pi < 0$) as an alternative hypothesis instead of $\pi \neq 0$ (see step 2 in Estimator 1).

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

Lemma 1 states that there can be significant distortions due to the pre-testing step in the suggested procedure that ultimately could lead to non-valid inference. There are two sources of first-order bias introduced by this pre-testing procedure, as we make it clear in the proof of lemma 1. The first is that this pre-test leads to an overestimation of the first-stage coefficient in the group that does not contain any compliers. This will tend to shrink the LATE estimator (in which the overall first-stage estimator appears in the denominator) towards 0. The second source of first-order bias comes from the fact that in the group without compliers ($G = 0$), we end up comparing always-takers with never-takers once we condition on the estimated first-stage $\hat{\pi}^0$ being larger than a threshold. This is not an issue when the expected outcomes of always-takers and never-takers are the same, as the effect will average out. However there is no reason for these expected outcomes to coincide. When they differ, their comparison leads to the introduction of an additional downward or upward bias first-order bias depending on whether the expected outcome of always-takers is larger or smaller (downward bias) than the one of never-takers.

We illustrate this bias using a Monte-Carlo simulation. Half the sample has zero compliance but with large variance, while the other half has a strong compliance rate. The details of the DGP can be found in 5. In this setting there is a significant chance to include zero compliance groups selectively, when not using sample splitting. Table I 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 "naive" 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 naive version of the Test-and-Select estimator exhibits a clear bias, which is ultimately detrimental to the coverage of its associated $95\%-$confidence interval that fail to cover at their nominal rate (0.828).

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

Algorithm 2 Test-and-Select

1: Divide the sample in two equally sized random subsamples $\mathcal{I}_1$ and $\mathcal{I}_2$, stratifying the random split by G.
2: In subsample $\mathcal{I}_1$: For each group defined by G, do a one-sided t-test on the first-stage coefficient π^g, at a pre-specified level α, testing the null $\pi^g = 0$ against the alternative $\pi > 0$ (or $\pi < 0$).
3: In subsample $\mathcal{I}_2$: Keep only the groups for which we rejected the previous test *in sample* $\mathcal{I}_1$.
4: Compute the usual Wald estimator on this selected subsample of $\mathcal{I}_2$.
5: Repeat steps 2 to 4 reversing the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ (cross-fitting).
6: Take the average of the estimators obtained in step 4 within $\mathcal{I}_1$ and $\mathcal{I}_2$.

The important data-splitting step happens in Steps 2 and 3 in the procedure of Estimator 2. We split the data in half randomly, and then use the first half to test for which groups have no compliers and the second half to compute the Wald estimator without those groups determined in the first half. The cross-fitting step then exchanges the roles of the two splits.

Our proposed methodology that associates the Test-and-Select procedure with sample splitting and (2-fold) cross-fitting yields an unbiased estimator, and nominal coverage (0.948) for the DGP presented in Table I.

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

TABLE I

PRE-TEST BIAS, AND THE USE OF CROSS-FITTING

	2SLS	Test-and-Select (with 2-fold-CF)	Test-and-select (without CF)
Mean Bias	0.02	0.004	-0.13
Median Bias	0.001	0.002	-0.10
Coverage	0.950	0.948	0.828
Confidence interval length	2.00	0.60	0.55

Note: This table presents the results of a simulation using the DGP described in appendix 4, with a number of groups of 30 with around 33 observations per group. In rows, we report the mean and median bias (with respect to the LATE parameter), the coverage rate of 95%-confidence intervals and the confidence interval length. 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 "naive" version of our methodology that would test, select and estimate the LATE in the same sample without any sample split.

3. THEORETICAL RESULTS

Throughout this section, we will consider a framework with two iid samples: a *test* sample (denoted $\mathcal{I}_T$) used to t-test on group-specific first-stage coefficients, and an *estimation* sample (denoted $\mathcal{I}_E$) used to compute the resulting estimator with the selection rule induced by the test results in $\mathcal{I}_T$. These 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. We will use the notation $n \to \infty$ to describe an asymptotic in both n_E and n_T simultaneously. At the end of the section, we will consider the use of cross-fitting, where the roles of $\mathcal{I}_T$ and $\mathcal{I}_E$ are reversed and the estimators are then averaged, as an attempt to mitigate the loss of precision induced by sample splitting.

The study of the properties of our suggested estimator is divided into two parts. First, we 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 introduce groups with a "local-to-zero" (or "weak") share of compliers. Following Staiger and Stock (1997), this means that the share of compliers in those groups decreases at a $1/\sqrt{n}$ rate, the same order of magnitude as sampling variation. This modeling choice better approximates finite-sample behavior of the estimator, because it allows for imperfect selection of groups

with non-zero shares of compliers.[11] As previously described, the population is partitioned by a grouping variable G. As before, we denote by π^g the share of compliers in group $G = g$. Additionally, we denote by $\mathcal{G}$ the support of G. In order to distinguish groups with "strong", "weak" and zero shares of compliers, we 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 work under the assumption that there are only two types of groups: the ones without any compliers, and the ones with a strong first-stage (i.e., a share of compliers well separated from 0).

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

Let $S \in \{0,1\}^{|\mathcal{G}|}$ denote an arbitrary selection vector, where $S_g = 1$ indicates that group $G = g$ is selected. The selected estimator is then given by:

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

[11] 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. In our framework we keep the share that each group g represents in the population stable with respect to the sample size.

In words, $\hat{\tau}(S)$ is the Wald estimator on the subsample $\{i : S_{G_i} = 1\}$, which is the subsample designated by S. For example, for $|\mathcal{G}| = 2$,

$$\hat{\tau}(S) = \begin{cases} \widehat{\text{Wald}} & \text{if} \quad S = (1,1) \\ \widehat{\text{Wald}}^{G=1} & \text{if} \quad S = (1,0) \\ \widehat{\text{Wald}}^{G=0} & \text{if} \quad S = (0,1) \end{cases}$$

where $\widehat{\text{Wald}}^{G=g}$ denotes the Wald estimator computed on the observations with $G = g$. We similarly denote $\hat{\tau}_E(S)$ for any arbitrary selection vector, the Wald estimator on the subsample selected by S *on the split dataset* $\mathcal{I}_E$.

The selection vector S of interest used in our proposed procedure is determined through group-by-group t-tests in the test sample $\mathcal{I}_T$.[12] We will denote it by $\hat{S}_T$, where the hat and subscript T indicate that this vector comes from an estimation step in sample $\mathcal{I}_T$. The estimator that is used in our method is then:[13]

$$\hat{\tau}_E \equiv \hat{\tau}_E\left(\hat{S}_T\right)$$

which corresponds to the Wald estimator computed on sample $\mathcal{I}_E$ for the groups selected on sample sample $\mathcal{I}_T$ by the group-by-group t-tests.

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

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

REMARK 1: It would be possible to decrease the threshold of the t-test at an appropriate rate in such a way that the probability to exclude groups with no first-stages goes to 1 as the sample size goes to infinity. we do not consider this type of testing for our selection

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

[13] Later in this section, we consider the use of cross-fitting, leading to the use of the "symmetric" estimator where the roles of the two samples are inverted.

procedure, because the resulting asymptotic approximation would likely not reflect accurately what happens in finite samples where the likelihood of keeping groups with zero first-stages would remain positive.

Lemma 2 shows that groups with strong first stages will always be selected asymptotically. To study the asymptotic distribution of $\hat{\tau}_E(S)$ when both the test and estimation sample sizes (n_T and n_E) tend to infinity, we can therefore look at any vector S which selects at least all groups with strong first stages. We denote this subset of all selection vectors that never exclude groups with strong first-stages by $\mathcal{S}_{strong} \subset \{0,1\}^{|\mathcal{G}|}$. Formally, for any $\tilde{S} \in \mathcal{S}_{strong}$, we have: $\forall g \in \mathcal{G}_S,\ \tilde{S}_g = 1$.

We then have the following proposition:

PROPOSITION 3: *Under assumptions 1, 2, and $E[Y^2] < \infty$, we have:*

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

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

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

where $\hat{V}^{\hat{\tau}_E(S)}$ is a consistent estimator of the asymptotic variance of $\hat{\tau}_E(S)$, and $q_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the $\mathcal{N}(0,1)$ distribution. Those confidence intervals are asymptotically valid by proposition 3.1, i.e.:

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

The following corollary states that such intervals have asymptotically valid *unconditional* coverage for the LATE. It also states that when the selection S is such that the asymptotic variance of the resulting estimator is strictly lower than the one of the TSLS estimator

(inequality case of proposition 3.2), then the length of a confidence interval conditional on such an S is going to be lower than usual CIs based on the TSLS estimator with probability going to 1 as n goes to infinity, reflecting the gains in terms of inference. Notice that the asymptotic study of confidence intervals lengths requires to rescale CIs by $\sqrt{n_E}$ to allow for a meaningful comparison.[14]

COROLLARY 4: *Under assumptions 1 and 2, if $E[Y^2] < \infty$ and S is such that we are in the inequality case of proposition 3.2, then the estimators $\hat{\tau}_E(S)$ and $\hat{\tau}_E^{TSLS}$ (the TS estimator conditional on S and TSLS estimator computed in split $\mathcal{I}_E$) are such that:*

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

Moreover, we have that:

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

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

Proposition 3 and corollary 4 show that, under assumption 2 ruling out the presence of subpopulations with weak first-stages, our procedure dominates the usual 2SLS/Wald estimator for estimation of and inference on the LATE parameter. However, the use of sample splitting is key to deriving those results, as it allows us to consider the selection process and the estimation as independent. Additionally, the variance comparison made in proposition 3.2 between our TS procedure and the 2SLS estimator is based on a comparison of asymptotic variances, while the second statement of corollary 4 assumes that the sample size used for estimation are identical when implementing our TS strategy or the usual TSLS estimation approach. However, given the sample splitting step inherent to our methodology, a fair comparison between the inference derived from the 2SLS approach and our proposed strategy has to take into account the reduction in sample size in the our approach. The reduced sample size lowers the gains in asymptotic variance. Consider again the same experiment as in the introduction, with a 10% compliance rate in the whole population, but

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

where compliers are all concentrated in half of the total population. The variance of the estimator would be halved compared to the variance of the usual TSLS estimator,[15] if the researcher has additional data, such as a pilot sample, which they can use to test and restrict to the complier population. This happens because compliance rate appears squared, such the halving of the sample size, which doubles the variance, is compensated by the higher compliance, which divides the variance by four. However, in general, such data is not available. In this case, the researcher will need to (randomly) split the sample into two sub-samples to implement our methodology, which reduces the size of the estimation sample in comparison to the usual TSLS estimation. Suppose they implement a 20%-80% split to create a test and estimation sample.[16] Then instead of dividing the size of the estimation sample by two (post-selection), they 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. Therefore, 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, which in the worst cases could to an increase in variance.

Cross-fitting The example above shows that the sample splitting step is not innocuous for precision. It is however a key step of our approach as it allows to make the testing-selecting and estimation steps independent. As shown in lemma 1 and illustrated in Table I, our procedure yields a biased estimator in the absence of sample splitting.

A possible solution to this problem consists in using both splits of the sample for both the testing-selecting and estimation steps by reversing their roles. This 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 $\mathcal{I}_2$. They then construct 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

[15] This is assuming homoscedasticity in order to simplify the computations for illustrative purposes, see equation 1.

[16] 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.

test sample and $\mathcal{I}_1$ as the estimation sample (see the description of our procedure in section 2, Estimator 2). This way, the entire sample is used for estimation, ideally recovering some of the losses from the sample-splitting step. 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. Therefore averaging those estimators can yield an estimator with the same variance than a hypothetical one constructed using the full sample, with an additional independent test sample used for selection. The following lemma establishes that a sufficient condition for such gains in variance to be restored is that the two cross-fit estimators are independent one from another.

LEMMA 5—Independence of cross-fit estimators: *Under assumptions 1 and 2, two estimators constructed following our suggested procedure and reversing the roles of two independent samples $\mathcal{I}_1$ and $\mathcal{I}_2$ are asymptotically independent one from another.*

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

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

where the first equality uses the independence between the limiting distributions of $\hat{\tau}_1$ and $\hat{\tau}_2$ demonstrated in lemma 5. Our cross-fitted TS estimator $(\hat{\tau}_1+\hat{\tau}_2)/2$ therefore has an asymptotic variance that is half the one of an estimator computed on a single split. Additionally, the sample splitting step results in a loss of a factor $\sqrt{2}$ in the speed of convergence (compared to the speed of convergence of an hypothetical TS estimator that could be computed on the whole sample of size n). The following theorem states that the gain in asymptotic variance described above exactly compensate the precision loss due to the sample split.

THEOREM 6—Efficiency gains of the cross-fitted estimator: *Let $\hat{\tau}_{CF}$ denote the average of the test-and-select estimators built in each folds of the data (using the other folds for the testing step). Then*

$$\sqrt{n}\cdot\frac{\hat{\tau}_{CF}-LATE}{\sqrt{V^{\hat{\tau}_{CF}}}}\xrightarrow[n\to\infty]{d}\mathcal{N}(0,1),\quad \textit{with } V^{\hat{\tau}_{CF}}\leq V^{TSLS}$$

where $V^{\hat{\tau}_{CF}}$ denotes the asymptotic variance of $\hat{\tau}_{CF}$, which can be consistently estimated based on the estimators of the asymptotic variance of fold-specific estimators.

PROOF: The asymptotic normality and the inequality on asymptotic variances is a direct consequence of Proposition 3 and Lemma 5, as already discussed above. Consistent estimation of $V^{\hat{\tau}_{CF}}$ merely results from the fact that the fold-specific variance estimators are constructed in independent samples, and the continuous mapping theorem. *Q.E.D.*

These results are encouraging as they suggest that *asymptotically* there are indeed gains in precision from testing and selecting a sub-sample with statistically significant first-stages. However, as already 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.[17] In particular, our framework so far ruled out the possibility to wrongly exclude groups with some compliers, because, 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 behavior in finite samples where groups with small shares of compliers might be wrongly excluded by the selection procedure. Therefore, we need to extend our framework in order to account for such cases.

3.2. *Asymptotic results with "weak" first-stages*

We now introduce groups with local-to-zero first-stages. Those groups have a share of compliers that goes as $1/\sqrt{n}$, so that a t-test will not systematically conclude that the first-stage coefficient is different from zero even when thesample size goes to infinity.

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

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

All values of g for which first-stages are weak are gathered in $\mathcal{G}_W$.
Additionally, for any group g, the share of observations and the group-LATE are fixed (they

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

don't vary with n). Formally:

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

We maintain the assumption of a strong first-stage overall (see assumption 1):

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

where c is a constant that does not depend on n. In other words, we still assume that there are some groups with strong first-stages in the population, such that the LATE in our setting is not weakly identified.

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

LEMMA 7—Asymptotic distribution of the selection procedure with some weak group first-stages: *Under assumptions 1 and 3, and if* $E[|Y^{2+\delta}|] < \infty$ *(for some* $\delta > 0$*), as the test sample size* n_T *goes to infinity, the probability to select groups with* 0 *first stages goes to* α *(the level of the* t*-test used), the probability to select groups with strong first-stages goes to 1, and the probability to select groups with weak ("local-to-zero") first-stages goes to values in the* $[\alpha,\ 1)$ *range, depending on the value of the localization parameter* H^g.

As in lemma 2, lemma 7 above justifies that when studying the asymptotic distribution of $\hat{\tau}_E(S)$ as both the test and estimation sample size tend to infinity, we only consider selection vectors S that satisfy: $\forall g \in \mathcal{G}_S,\ S_g = 1$ (where S_g denotes the g-th term of vector S). This is because asymptotically, we never wrongly exclude groups with strong first-stages. However, this is not the case for groups with weak first-stages, who are excluded with a non-zero probability (even asymptotically) despite their non-zero share of compliers. In the previous subsection 3.1 and its associated proposition 3, we showed that in the absence of groups with weak first-stages, our estimator can yield precision gains without introducing any first-order bias. The following proposition (the analog to proposition 3) shows that this is no longer true in the presence of some groups with weak first-stages.

PROPOSITION 8: *Under assumptions 1 and 3, and if* $E[|Y^{2+\delta}|] < \infty$ *(for some* $\delta > 0$*), we have:*

1. $\forall S \in \mathcal{S}_{strong}$,[18] $\sqrt{n_E}\,(\hat{\tau}_E(S) - LATE) \xrightarrow{d} \mathcal{N}(B(S), V^S))$.
2. $B(S) \propto |LATE^{W(S)} - LATE|$, *where* $LATE^{W(S)}$ *denotes the average treatment effect among compliers within groups with weak first-stages that are wrongly dropped by selection procedure* S.
3. $B(S) \neq 0$ *if* $\exists j$ *s.t.* $\{S_j = 0 \cap j \in \mathcal{G}_W\}$ *and* $LATE^{W(S)} \neq LATE$.

Without any further assumptions on treatment effect heterogeneity, the above proposition shows that our proposed estimator will systematically be first-order biased in the presence of groups with weak first-stages. Indeed, the probability of wrongly excluding these groups does not go to zero asymptotically (see lemma 7) and proposition 8.3 shows that in the presence of such exclusion errors, the first-order bias of our procedure is non-zero. The source of this bias is that the LATE of groups with a weak share of compliers might differ from the LATE of groups that are kept for the estimation step. If we bundle all groups with a weak first-stage in a single group $G = 2$, and all groups with a strong first-stage in $G = 1$, the asymptotic bias (conditional on the event that group $G = 2$ is dropped from the estimation step) would take the following form:

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

where $\pi \equiv \mathrm{P}[D(1) > D(0)]$ is the share of compliers in the population, $LATE^g \equiv E[Y(1) - Y(0) \mid D(1) > D(0),\ G = g]$ is the LATE in group $G = g$ and H^2 is the localization parameter for the first stage in group $G = 2$. This expression also shows why it "only" creates a first-order bias. Indeed, the share of compliers with $G = 2$ among all compliers decreases at a $\sqrt{n}$-rate under assumption 3. Therefore, even once rescaled by $\sqrt{n}$, the bias (with respect to the LATE parameter) remains bounded as long as the treatment effect heterogeneity term $\left(LATE^1 - LATE^2\right)$ is bounded.

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

In order to better grasp the nature of the first-order bias of our estimator, corollary 9 provides sufficient conditions on treatment effect heterogeneity for our estimator to remain first-order unbiased.

COROLLARY 9: *Under assumptions 1 and 3, $E[|Y^{2+\delta}|] < \infty$ (for some $\delta > 0$), and homogeneous treatment effects, we have that $\hat{\tau}_E(S)$ is first-order unbiased and asymptotically normal, i.e.:*

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

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

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

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

Assuming homogeneous treatment effect is generally not a not realistic assumption, and rather in opposition to the spirit of the LATE literature. On the other hand, vanishing treatment heterogeneity might be a realistic assumption to describe the data-generating processes studied in applied economics and social sciences in general. This type of restriction can be motivated by the usual difficulties faced by researchers in detecting treatment effect heterogeneity in empirical research, given the usual sample sizes at their disposal. Coussens and Spiess (2021) also studied the properties of their proposed estimator under the related but stronger assumption that average treatment effect in general is of the order of magnitude of $1/\sqrt{n}$. We will therefore now study the properties of our estimator under this restrictions on treatment effect heterogeneity.

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

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

The following theorem establishes the key results on the asymptotic distribution of our estimator under this assumption.

THEOREM 10: *Under assumptions 1, 3, 4, and if* $E[|Y^{2+\delta}|] < \infty$ *(for some* $\delta > 0$*), we have:*

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

REMARK 2: Notice that the results presented in the theorem above would hold under the less stringent assumption of vanishing treatment effect heterogeneity:

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

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

Theorem 10 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 4) remain following the same reasoning. Compared to alternatives such as Coussens and Spiess (2021) and Huntington-Klein (2020), our procedure is exempt of any first-order bias under the restriction on treatment effect heterogeneity made in assumption 4, as shown in lemma 15 and its proof in appendix 2, which also contains a proof of the bias of Coussens and Spiess (2021) procedure in our framework.[19]

The intuition behind the better 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 *unsquared* share of compliers). 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

[19] 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.

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. This results in the conditional LATEs of these groups being weighted less than they should to match with the overall LATE parameter. However, 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. Therefore, the deviation from the LATE in our case is less important than in Coussens and Spiess (2021), and the restriction on heterogeneity made in assumption 4 is sufficient to rule out any first-order bias. This discrepancy in the behavior of our estimator compared to the one of Coussens and Spiess (2021) as highlights two important points:

1. the heterogeneity restriction made in assumption 4 is not 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).

Motivating assumption 4 We conclude our discussion by providing an argument that assumption 4 can be justified in empirical settings. When implementing an encouragement desigb, it is 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 $\text{Var}[Y(0)] = \text{Var}[Y(1)] = \sigma^2$, and q_x is the x^{th}-quantile of a $\mathcal{N}(0,1)$. Therefore, in studies designed based on power analyses, we have by design: $e^* = O(n^{-1/2})$. The true effect (and treatment effect heterogeneity) can be larger than e^*, in which case our study will systematically detect the effect of the policy (and its heterogeneity). But as has been document using meta-analyses in for instance Ioannidis et al. (2017), social sciences (and economics in particular) are generally *under*-powered rather than over-powered. Experimenters in social sciences certainly do not detect 100% of the time significant effects (and even less often treatment effect *heterogeneity*). It is therefore plausible 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. SIMULATIONS

This section presents a simulation study that compares the performance of the various estimators mentioned above: the standard 2SLS estimator, 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 $\text{E}\left[Y(1) - Y(0) | D(1) > D(0),\ G = g\right]$ and compliance rates π^g.

DGP parameters We follow Coussens and Spiess (2021) in using the threshold crossing model representation (Vytlacil, 2002), but differ importantly by introducing parameters which control the correlation between conditional LATEs $\text{E}\left[Y(1) - Y(0) | D(1) > D(0),\ G = g\right]$ and compliance rates π^g. These are important because it is precisely the condition leading to a first-order bias in our estimation strategies, and is therefore important to illustrate when these methods fail. The details of the DGP are presented in Appendix 4.

The Monte Carlo simulations are ultimately 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
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

For both DGPs we present below, most of these parameters are fixed:

$$\text{DGP1} \equiv \left(N = 1000, J = 10, S_{AT} = S_{NT} = \frac{0.75}{2}, \rho_{\delta\varepsilon} = 0.5, \alpha = 0.5\right)$$

In other words, we are studying are considering a sample size of $1,000$ observations, partitioned into 10 groups, 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 is significant 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 for our estimator or these alternatives.

In the simulations below, we focus on two important aspects. The first is the presence of groups without compliers and how well it is predicted by the covariates, controlled by the σ_η^2 parameter. This is the most important parameter because it controls whether the methods can be expected to result in any improvements. The second is the strength of the overall treatment effect heterogeneity, controlled by the β parameter. This parameter controls the bias the methods will exhibit.

We will therefore present 2 classes of DGPs. 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. On the other hand, the second DGP (DGP2) has no groups with 0 compliers, but 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 DGP illustrates the robustness of the different

methodologies to levels of treatment effect heterogeneity in an adverse DGP the methods are not expected to generate any gains in precision.

DGP1: a "best-case scenario" The parameters for the first DGP are the following:

$$\text{DGP1} \equiv \Big(\sigma_\eta = 0.01, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\}\Big)$$

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

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

with an overall compliance rate of 25%. DGP1 is an ideal application for our method, because 60% of groups have no compliers and the other groups have large compliance rates.

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):

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

where q_x represents the quantile function of a normal distribution. 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. A key result in section 3 is that our estimator exhibits no first-order bias of our estimator when treatment effect heterogeneity is of the order of $1/\sqrt{n}$, which highlights its robustness to moderate treatment effect heterogeneity. As the MDE is of order $1/\sqrt{n}$, it is a natural to quantify the level of heterogeneity, and therefore the likelihood of bias.

Figure 1 presents the bias, length and coverage of 95%-confidence intervals, 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. As expected, estimators based on interacted of weighted instruments display much larger amounts of bias at any level of treatment effect heterogeneity (except zero). Panel 1c shows that translates

to poor coverage rates for these estimation strategies, while ours covers at the nominal level for any amount of treatment effect heterogeneity.

Moreover, panel 1b highlights the large decrease in confidence interval length (for all alternative estimation methods) compared to the standard 2SLS. For this DGP, our estimation procedure generates significant gains compared to the standard 2SLS estimator, but as expected, the other estimators generate even larger gains in precision. Panel 1d shows that in this case, our method has the lowest RMSE, showing that the lower bias outweighs the precision gains.

DGP2: introduction of "weak" compliance groups The parameters for the first DGP are the following:

$$\text{DGP2} \equiv \left(\sigma_\eta = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\}\right)$$

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

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

with an overall compliance of 25%. In contrast, to the first DGP, this DGP does not feature any groups without compliers. Instead, it introduces several "weak" compliance groups, which are (as studied in 3) the main source of bias for our estimator.

As before, we run a Monte-Carlo simulation with 10,000 repetitions. The results are shown in the panels of figure 2 as a function of the MDE. Compared to what we observed in DGP1, panel 2a highlights a much larger bias of our procedure, that grows as treatment effect heterogeneity increases, as expected from our theoretical results from section 3. However, it remains significantly lower than the bias of the alternatives estimators. As before, all these alternatives have lower variance than our method, yielding shorter confidence intervals, as shown in panel 2b. However, because of the bias, these shorter confidence intervals yield significantly worse coverage properties than our estimator as shown in panel 2c.

As demonstrated in section 3 our estimator remains unbiased *to first-order* when treatment effect heterogeneity is moderate, but such a property does not guarantee unbiasedness in finite samples which can clearly be observed in panel 2a. However it allows for valid inference as long as treatment effect heterogeneity remains moderate, which is precisely what

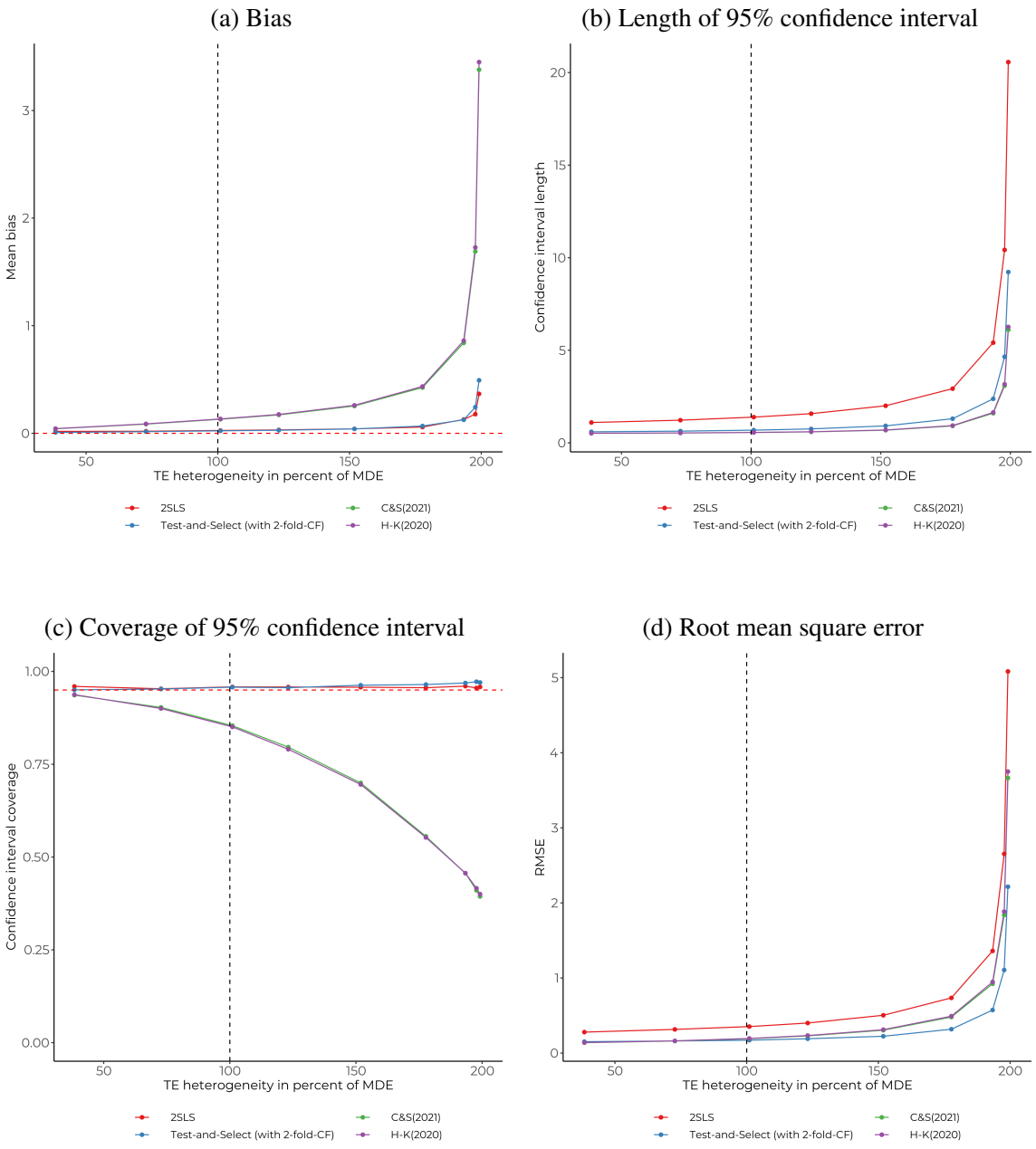

FIGURE 1.—Comparison of estimators with varying treatment effect heterogeneity for DGP1. This panel shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP1, described in the text. Four different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select 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.

can be seen in panel 2c. As treatment effect heterogeneity grows, the coverage of our estimator remains at its nominal level at least up to a heterogeneity of the order of the MDE, and only starts deviating slowly after. On the contrary, the alternative estimators deviate from nominal coverage rapidly. Lastly, panel 2d shows that the ordering of estimators in terms of RMSE is ambiguous, depending on the level of treatment effect heterogeneity. The RMSE criterion is less interest to us in this paper as the goal is to provide an estimator with little bias with valid inference.

5. APPLICATION TO A NATURAL EXPERIMENT ON COMPULSORY SCHOOLING LAWS

In this section, we apply our proposed methodology to the return to schooling literature, which studies whether an extra year of schooling affects later life outcomes such as wages. In particular, we look at a well-studied natural experiment which uses variation in compulsory schooling laws across states and across time as an instrument for years of schooling (see for instance Acemoglu and Angrist (2006), Oreopoulos (2006), Stephens and Yang (2014)).

We analyze public-use Census-data provided by Stephens and Yang (2014). We use the interaction between demographic controls (ethnicity $\times$ sex) $\times$ US census division $\times$ survey year (1960, 1970, 1980) as our main covariate and then exclude from the sample the cells without variation in compulsory schooling laws. We discretize the instrument (to having more than 7 years of remaining compulsory years of schooling at age 6) and the treatment (to having at least some high-school education). More details on the exact restrictions can be found in 5.

In Table II, we report the probability that a G-cell is selected by demographic group. As expected from the observation of Stephens and Yang (2014) that ethnic minorities tend to react less to the compulsory schooling laws instrument, our test-and-select procedure based on a one-sided t-test on the first stage coefficient tends to select groups of white individuals.

Table III compares the result of the standard 2SLS approach and our proposed Test-and-Select estimator with two different levels for the pre-test (0.05 and 0.01), *without* covariate controls. The point estimate of the 2SLS estimator (1.861) differs from the ones of the TS estimators (1.470 and 1.302), but the the standard errors associated to the TS estimators are reduced by around 12%.

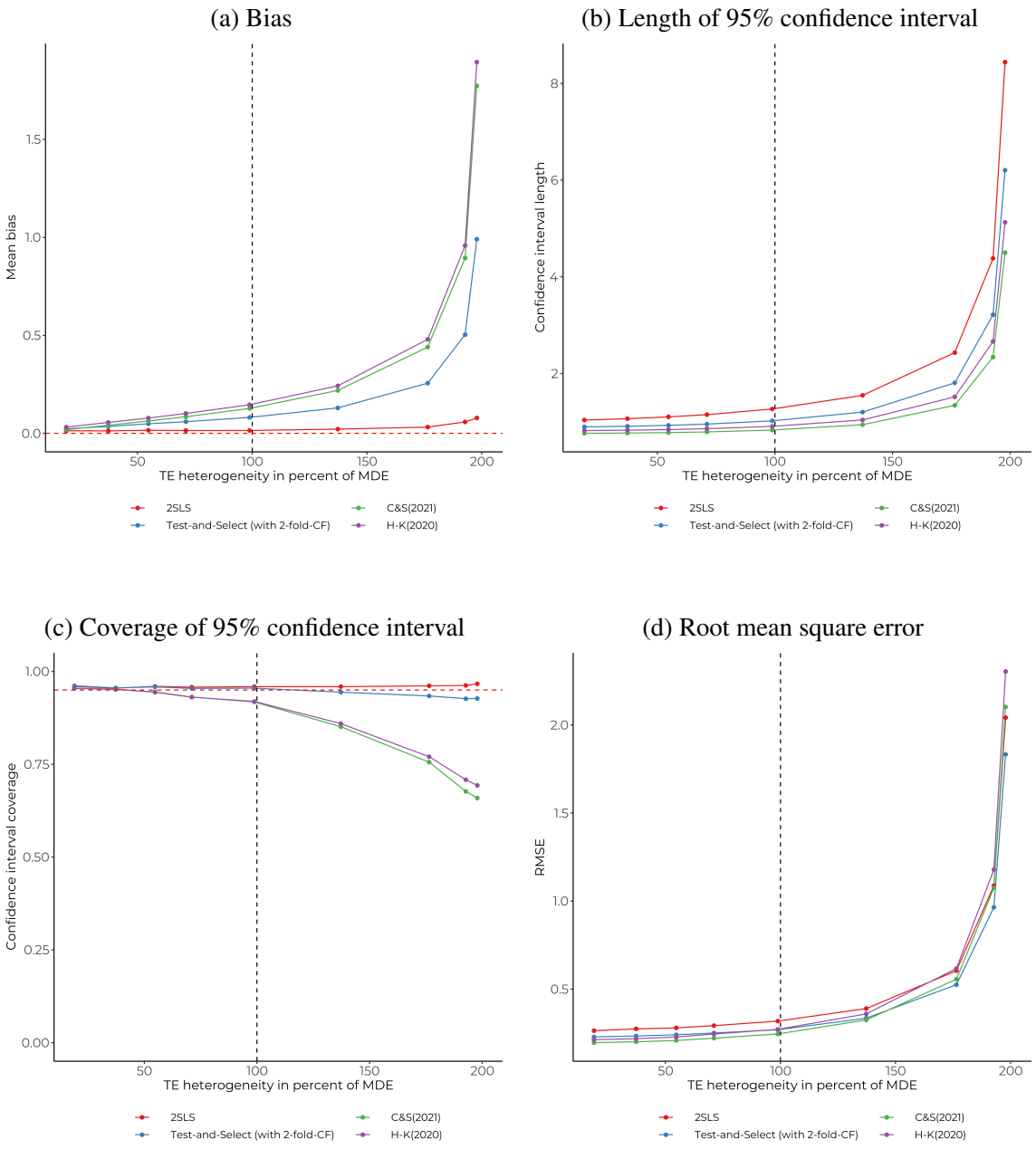

FIGURE 2.—Comparison of estimators with varying treatment effect heterogeneity for DGP2. This panel shows the results of a 10,000 repetitions of a Monte-Carlo simulation of DGP2, described in the text. Four different estimators are considered: the standard 2SLS estimator in red, our proposed Test-and-Select 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.

TABLE II

SELECTION PROBABILITY OF G-CELLS, BY DEMOGRAPHIC GROUP

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

Note: G is a partition of the population along demographic controls (ethnicity $\times$ sex) $\times$ US census division $\times$ survey year (1960, 1970, 1980), which defines 108 cells, 72 of which are kept in the analysis. This table presents the probability that a cell involving a given demographic group is dropped from the estimation sample once we select based on a one-sided t-test with level 0.05 (first column) or 0.01 (second column).

TABLE III

COMPARISON OF ESTIMATION METHODS WITHOUT COVARIATE CONTROLS

	2SLS	Test-and-Select (0.05)	Test-and-Select (0.01)
	1.861	1.470	1.302
	(0.365)	(0.320)	(0.312)
	[1.145, 2.578]	[0.843, 2.098]	[0.691, 1.913]
First-stage coef.	0.523	0.518	0.513
%. sample drop.	0	28.6	32.7
N	171 096	122 150	115 159

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

Simply comparing the results of our estimator in table III with the Interacted IV approach would be uninformative because it controls linearly for covariates in the second-stage. However, controlling for G linearly can result in estimators (2SLS and TS) not targeting the LATE estimand anymore. [20] This is highly likely to be the case in this context as the instrumental variable conditions are more likely to hold conditionally. In order to ensure

[20] In fact, sometimes they could even target a parameter that is a non-convex weighted average of conditional LATEs (Słoczyński, 2022). To be precise, under both assumptions of (i) monotonicity and (ii) complete randomization of Z (i.e., Z is independent of the potential outcomes *and* of G), controlling linearly for G does not change the targeted parameter (compared to the situation where we do not control at all for G). However, in natural or stratified experiments, it could very well be that the distribution of Z varies across along G. In this case (where

TABLE IV

COMPARISON OF ESTIMATION METHODS WITH COVARIATE CONTROLS

	Frölich (2007)	TS (0.05) & Frölich (2007)	TS (0.01) & Frölich (2007)	Interacted IV	Interacted IV & Select (0.05)
	0.907	0.791	0.776	1.348	1.149
	(1.16)	(0.604)	(0.576)	(0.182)	(0.149)
	[−1.367, 3.181]	[−0.392, 1.975]	[−0.353, 1.905]	[0.991, 1.705]	[0.858, 1.441]
First-stage coef.	0.064	0.094	0.097	0.088	0.103
%. sample drop.	0	28.6	32.7	0	28.6
N	171 096	122 150	115 159	171 096	122 150

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

that our estimator still targets the LATE (even before applying our TS methodology), we use an estimator suggested by Frölich (2007), which controls for G non-parametrically.[21]

We report the results of this comparison in table IV. As expected, the resulting point estimates differ significantly from the ones previously documented in table III. Still, in this application, our procedure yields considerable variance gains from 1.16 to 0.604, a reduction by around 48% of the standard errors. The variance of the TS estimator remains larger than the one of the interacted IV estimators in this application. However the point estimates of these differ significantly more from the standard Frölich point estimate, suggesting larger bias and incorrect inference.

6. EXTENSIONS

In this section, we present some possible extensions. We start by discussing how to apply our methodology when covariates are high-dimensional, including the case of continuous covariates. We then discuss a possible re-weighting strategy, which weights groups by their probability of having compliers. Finally, we consider a breakdown analysis which provides a measure of robustness to treatment effect heterogeneity using worst-case bias bounds.

---

the second assumption does not hold anymore), the linear control for G yields an estimator of a convex weighted average of conditional LATEs, *yet* with different weights than the natural ones.

[21] At this stage, this paper does not include a formal discussion of the variance gains of the TS procedure when coupled with the Frölich estimator, but we believe that our results continue to hold.

High-dimensional groups Assuming covariates can define groups with weak/0 share of compliers is arguably more credible when covariates are high dimensional (e.g., when there is a large number of covariates, interactions between covariates, continuous covariates etc.). We now sketch how to adapt our procedure to this setting. We maintain the assumption of strong identification overall, i.e. $\pi > 0$. We then follow Chernozhukov et al. (2021) with slight modifications:

1. Build a flexible prediction of conditional compliance rate $s(X) \equiv \mathrm{E}[D(1) - D(0)|X]$, denoted $\hat{s}(X)$. There is no assumption on the rate of convergence of $\hat{s}(X)$. The goal is merely that $\hat{s}(X)$ contains some signal for the true $s(X)$.
2. Define $\bar{G}$ (a fixed number) groups based on quantiles of $\hat{s}(X)$, and partition the population into $\bar{G}$ groups: $\{\mathbb{I}_{\hat{s}(x)\in[q_{g-1},q_g]}\}_{g\in\{1,...,\bar{G}\}}$

After this procedure, we are back in a situation with discrete covariates where our method can be applied.

The direct approach as described above suffers from a potential bias, related to the fact that $\hat{s}(X)$ is created using the same data that is afterwards used for estimation of the LATE. Chernozhukov et al. (2021) solve this problem by introducing a variational approach, which relies on repeated data-splits. This approach does not work in our case since it does not allow classifying each observation into one group, as the latter changes with each split. We therefore suggest a single data-splitting and cross-fitting procedure, where the $\hat{s}(X)$ in the first split is computed using data from the second split and vice versa. It is important for researchers to commit to a single split to avoid p-hacking. This procedure does not entirely solve the potential correlation problem, but in large enough samples this should be negligible.

Re-weighting strategy Instead of taking a binary decision to either drop or include groups in the estimation sample, an alternative could be to re-weight groups based on their probability to have a zero 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. Our main results might still hold for such weighted estimators since (asymptotically) groups with strong first-stages would have $\Pr[\text{sh. of compliers} = 0|g \in \mathcal{G}_S]$ that goes to 0, and therefore a weight that goes to 1 as in our proposed estimation strategy. Alternatively, it is possible that such this procedure could be 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).

Notice that this would still be distinct from Coussens and Spiess (2021) "weighted-IV" approach, as our weights would tend to 1 and be uniform among all groups with a strong first-stage. This way, we could still hope that changes in the targeted estimand remain negligible under restrictions on treatment heterogeneity of the type described in assumption 4, which is not the case for the "weighted-IV" approach (see lemma 15).

Breakdown analysis Instead of relying on an assumption of the type of assumption 4, we sketch a procedure to correct inferential statements on our estimator for worst-case bias under a sensitivity parameter corresponding to the maximal LATE heterogeneity across groups.

We proved in section 3 that the bias of our estimator with respect to 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.

As $LATE^1$and $LATE^2$ depend on the realization of the sample, in what follows we condition on the sample realization.[22] The quantity $\mathrm{P}[G = 2|D(1) > D(0)]$ can be estimated by 2SLS as suggested in Abadie (2003). Since by construction of $G = 2$, Z is a weak instrument for D in this subpopulation, $\mathrm{P}[G = 2|D(1) > D(0)]$ cannot be consistently estimated. However, an asymptotically valid $(1 - \alpha)$-confidence interval can be constructed using for instance inversion of an Anderson-Rubin statistic.

Suppose we construct 99%-confidence interval 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 $\mathrm{P}[G =$

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

$2|D(1) > D(0)]$, hence the *worst-case* bias can be obtained by replacing $\mathrm{P}[G = 2|D(1) > D(0)]$ with $\hat{UB}^P$. The bias is then determined by $\left(LATE^2 - LATE^1\right) \equiv M$, which we now take as the sensitivity parameter. For a given value of M, the worst-case bias of our proposed estimator for the LATE is given by $M \cdot \widehat{UB}^P$. We can construct a confidence interval that is (asymptotically) valid with a 95% coverage for the overall LATE parameter by widening 96%-confidence interval around $LATE^1$, the effect among compliers in the population selected by our procedure, by $\pm M \cdot \widehat{UB}^P$ original confidence intervals.[23].

This procedure then allows to conduct a "breakdown" analysis which consists in determining how the confidence intervals vary with M. In particular, it allows determinination of the value of M, the LATE treatment heterogeneity, for which a threshold value, for instance zero, ceases to be included in the confidence interval. This provides a measure of robustness to treatment effect heterogeneity of our proposed estimation strategy.

7. CONCLUSION

In this paper, we study 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 first show that when the groups with no-compliance are *known*, the corresponding oracle estimator targets the same LATE estimand with significantly less variance. These properties extend to a feasible estimator that identifies zero compliance group by t-testing the first-stage coefficients as we consider standard asymptotic sequences in which compliance rates per group are either zero or fixed with n and well-separated from 0. These sequences are unlikely to provide a credible approximation for the finite-sample behavior of our estimator. We therefore next consider weak-IV-like asymptotic sequences in which some groups display local-to-zero compliance rates, which allows exclusion errors even asymptotically. We show that a restriction on treatment effect heterogeneity is a sufficient condition for our estimator to remain first-order unbiased with respect to the LATE under such asymptotic sequences. We discuss the validity of this restriction in applied work and compare the performance of our estimator to alterna-

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

tive procedures recently proposed in the literature, which exploit first-stage heterogeneity differently from us. The main takeaway from this discussion is that our estimator yields a different variance-bias trade-off than other proposed alternatives in the literature. It is less biased, because it is more robust to treatment effect heterogeneity, but at the cost of potentially smaller variance gains. Finally, we illustrate the potential gains our estimator brings in Monte-Carlo simulations and an application to a natural experiment in education.

The econometrics literature on the use of first-stage heterogeneity in LATE estimation is currently very active and with many promising research avenues. In a follow-up project (joint with X. D'Hautefœuille), we flip the perspective of this paper and study what imposing bounds on the LATE by group can yield in terms of precision gains. Another related research avenue we are planning on investigating in future research is how to incorporate heterogeneity in compliance when *designing* experiments with imperfect compliance.

APPENDIX A: PROOF OF MAIN RESULTS

PROOF OF THEOREM 10.1: Lemma 14 and proposition 8 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 $\sqrt{n_E}\left(\hat{\tau}_E(S) - LATE\right)$ 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 $\hat{\tau}_E(S)$ has naturally the same form as the one studied

in 3. Hence following the reasoning of the proofs of proposition 3.1 and 3.2—yet using appropriate central limit theorem for triangular arrays (Lindeberg-Feller CLT) instead of the standard CLT—we get:

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

Q.E.D.

PROOF OF THEOREM 10.2: The proof follows the exact same line of reasoning as in the proof of 3.3, yet making use of assumption 4 and its implication in theorem 10.1 to get the result. Indeed, the proof relies on the consistency of $\hat{\tau}_E(S)$ 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 10.1.

Q.E.D.

REFERENCES

ABADIE, ALBERTO (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113 (2), 231 – 263. [35]

ABADIE, ALBERTO, JIAYING GU, AND SHU SHEN (2022): "Instrumental Variable Estimation with First-Stage Heterogeneity," *Journal of Econometrics (forthcoming)*. [3, 5, 9, 26]

ACEMOGLU, D. AND J. ANGRIST (2006): "How Large are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," *NBER Macroeconomics Annual 2000*. [30]

ANGRIST, J., G. IMBENS, AND E. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*. [6, 7]

ANGRIST, J. AND A. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*. [3]

CHERNOZHUKOV, VICTOR, MERT DEMIRER, ESTHER DUFLO, AND IVÁN FERNÁNDEZ-VAL (2021): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," *arXiv*. [34]

CLAESKENS, G AND N HJORT (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*. [35]

COUSSENS, STEPHEN AND JANN SPIESS (2021): "Instrumental Variable Estimation with First-Stage Heterogeneity," *Working Paper*. [3, 5, 22, 23, 24, 25, 26, 29, 31, 35, 60, 61, 62]

FRÖLICH, MARKUS (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139 (1), 35 – 75, endogeneity, instruments and identification. [6, 33]

HONG, HAN AND DENIS NEKIPELOV (2010): "Semiparametric efficiency in nonlinear LATE models," *Working paper*. [6]

HUNTINGTON-KLEIN, NICK (2020): "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," *Journal of Causal Inference*. [3, 5, 23, 25, 26, 29, 31]

IMBENS, GUIDO AND DONALD RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. [6]

IMBENS, G. AND E. RUBIN (1997): "Bayesian Inference for Causal Effects in Randomized Experiments with Non-Compliance," *The Annals of Statistics*. [5]

IOANNIDIS, JOHN P. A., T. D. STANLEY, AND HRISTOS DOUCOULIAGOS (2017): "The Power of Bias in Economics," *The Economic Journal*. [25]

KENNEDY, EDWARD H. (2023): "Semiparametric doubly robust targeted double machine learning: a review," . [54]

KITAGAWA, T AND C. MURIS (2016): "Model averaging in semiparametric estimation of treatment effects," *Journal of Econometrics*. [35]

LEEB, H. AND B. PÖTSCHER (2005): "Model selection and inference-Facts and Fiction," *Econometric theory*. [6, 9, 19]

OREOPOULOS, P. (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*. [30]

RUBIN, E. (1998): "More powerful randomization-based p-values in double-blind trials with non-compliance," *Statistics in Medicine*. [6]

STAIGER, E. AND J. STOCK (1997): "Instrumental variables regressions with weak instruments," *Econometrica*. [4, 8, 12]

STEPHENS, M. AND D. YANG (2014): "Compulsory Education and the Benefits of Schooling," *American Economic Review*. [3, 30, 63]

SŁOCZYŃSKI, T. (2022): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *Working Paper*. [32]

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*. [25, 61]

ONLINE APPENDIX
"IMPROVING LATE ESTIMATION IN EXPERIMENTS WITH IMPERFECT COMPLIANCE"

YAGAN HAZARD
Collegio Carlo Alberto and ESOMAS, University of Turin

SIMON LÖWE

1. PROOFS OF MAIN LEMMAS AND PROPOSITIONS

PROOF OF PROPOSITION 3.1: We'll closely follow the proof of lemma 13, 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 use the following notation:

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

Notice that Propositions 3.1 and 3.2 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. $\{\forall g \in \mathcal{G}_S, S_g = 1\} \cap \{\exists g \in \mathcal{G}_0, S_g = 1\}$

Yagan Hazard: yagan.hazard@carloalberto.org
Simon Löwe : loewe.sim@gmail.com

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

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

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

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

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

$$\text{hence} \quad LATE = \frac{A-B}{C-D} \quad \text{and} \quad \widehat{\tau}_E(S) = \frac{\hat{A}-\hat{B}}{\hat{C}-\hat{D}}$$

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:

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

In exactly the same way as the proof of lemma 13, we have:

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

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

Therefore we get:

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

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

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,

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

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

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

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

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

where $p_{S_G} \equiv \Pr[S_G = 1]$. *Q.E.D.*

PROOF OF PROPOSITION 3.2: From lemma 13, and from proposition 3.1 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:

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

where the first equality follows from the law of total variance, and first and second inequalities follow from the fact that variances are always positive or null (in degenerate cases). Therefore, V^{TSLS} has been shown to be a linear combination (with positive coefficients) of terms greater or equal than the ones appearing in $V^{\hat{\tau}(S)}$, proving the proposition 3.2. *Q.E.D.*

PROOF OF PROPOSITION 3.3: In order to properly study the asymptotic distribution of $\hat{\tau}_E = \hat{\tau}(\hat{S}_T)$, 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 $\hat{S}_{T,g}$, the g^{th} line of vector $\hat{S}_T$, which is defined as follows:

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

where $n^g_{(T)}$ is the number of observations in group g in sample $\mathcal{I}_T$, $\hat{\pi}^g_{(T)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance of $\hat{\pi}^g_{(T)}$. Notice that $\hat{\pi}^g_{(T)}$ is asymptotically linear, as following lemma 11 we have:

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

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

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

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

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

where the expression of the influence function is given by:

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

The above display makes it clear that the $\psi_{\hat{\tau}_E,i}$'s of individuals within a given group g are dependent, as they all depend on $\hat{S}_{T,G}$, the selection indicator computed in the test sample $\mathcal{I}_T$. However, 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 $\psi_{\hat{\tau}_E,i}$'s, and we are back to the case studied in proposition 3.1 and 3.2. Now let us define:

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

where $V^{\hat{\tau}_E}$ is the asymptotic variance of $\hat{\tau}_E$. 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 3.1 we have that for $\hat{S}_T \in \mathcal{S}_{strong}$, $\hat{T}_E$ converges to a $\mathcal{N}(0,1)$. And by consistency of the t-test against any alternative well separated from 0, we have that all groups with strong first-stages are selected asymptotically, implying: $\forall S \notin \mathcal{S}_{strong},\ \mathbb{1}\{\hat{S}_T = S\} \xrightarrow[n\to\infty]{p} 0$, hence the second line of the above display (by continuous mapping theorem).

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

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

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

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

By Jensen inequality again, we have:

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

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

PROOF OF COROLLARY 4: Firstly, by proposition 3.1 we have that for any realization $\hat{S}$ of $S \in \mathcal{S}_{\text{strong}}$, one can build asymptotically valid *conditional* confidence intervals with coverage $(1-\alpha)$ in the usual way:

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

where $\hat{V}^{\hat{\tau}_E(S)}$ is a consistent estimator of the asymptotic variance of $\hat{\tau}_E(S)$, and $q_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the $\mathcal{N}(0,1)$ distribution. Those CIs are asymptotically valid by proposition 3.1, i.e.:

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

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

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

which is the second statement of corollary 4.

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

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

$\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha(S)]$ and $\sqrt{n_E}\cdot \mathrm{length}[CI_\alpha^{TSLS}]$ are entirely governed by and strictly increasing in $\hat{V}^{\hat{\tau}(S)}$ and $\hat{V}^{\mathrm{TSLS}}$ respectively. Let $\hat{V}^{\hat{\tau}(S)}$ and $\hat{V}^{\mathrm{TSLS}}$ be estimators that converge in probability to $V^{\hat{\tau}(S)}$ and V^{TSLS}, and we assumed that S was such that we were in the inequality case of proposition 3.2, i.e.,

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

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

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

and

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

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

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

Hence we have:

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

Q.E.D.

PROOF OF PROPOSITION 8: Lemma 7 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 $\hat{\tau}_E(S)$ is 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 3 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 3 that the asymptotic distribution of $\hat{\tau}(S)$ in such a setting will be centered on the LATE. Formally, let us defined:

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

By proposition 3 and the argument above, we have:

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

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

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

Recasting the problem in this way places it in the setting studied in lemma 14, which proves the result.

Q.E.D.

2. PROOFS OF LEMMAS

PROOF OF LEMMA 1: Let G be a binary covariate partitioning the population such that:

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

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

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

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

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

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

(denoted $\widehat{LATE}_1$) will be asymptotically normal and centered on $LATE_1$:

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

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

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

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

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

where $\mathcal{N}(0, LB = t, V_{\hat{\pi}^0})$ denote the distribution of a truncated normal distribution $\mathcal{N}(0, V_{\hat{\pi}^0})$ with lower bound t. This distribution is not centered at 0. Therefore, since $\widehat{ITT_1}$ does not go to 0, we already have that our first bias term A does not vanish anymore. This is a first source of first-order bias in the estimation of the LATE with this naive pre-testing procedure. This one is intuitive: as our pre-test tends to select cases in which we overestimate the share of compliers in group $G = 0$, we tend to overestimate the overall share of

compliers, and thus this shrinks the estimator towards 0.

However, there is potentially a second source of bias that comes from the non causal comparison between always-takers and never-takers in group $G = 0$. Indeed, since there are no compliers in this group, having a large first-stage in $G = 0$ necessarily means that there is an imbalance between the share of always takers and the share of never-takers in this subsample. If we do not condition on the size of the estimated first-stage coefficient $\hat{\pi}^0$, then we still have that those shares are balanced on average, and thus we have $\widehat{ITT^0} \xrightarrow[n\to\infty]{p} 0$ and, by Slutsky's lemma $\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \tilde{V}_0)$. However, once we condition on the estimated first-stage coefficient, the probability limit of $\widehat{ITT^0}$ and the limiting distribution of $\sqrt{n} \cdot \hat{\pi}^0 \cdot \widehat{ITT^0}$ are quite different. Indeed, we have:

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

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

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

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

Q.E.D.

PROOF OF LEMMA 5: As showed previously, if we denote by $\hat{\tau}_1$ the estimator constructed using the fold $\mathcal{I}_2$ as the test sample and $\mathcal{I}_1$ as the estimation sample, recall that

we can decompose it as follows:

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

where the expression of the influence function is given by:

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

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

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

where $n^g_{(2)}$ is the number of observations in group g in sample $\mathcal{I}_1$, $\hat{\pi}^g_{(2)}$ is the (difference in means) estimator of the first-stage in group g, and $\hat{\sigma}^{\pi^g}$ is a consistent estimator of the (asymptotic) variance of $\hat{\pi}^g_{(2)}$. Second, recall (from the proof of corollary 4 above) that:

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

The above formulas make it clear that the potential source of dependence between $\hat{\tau}_1$ and $\hat{\tau}_2$ lies in $\hat{S}_{g,(2)}$, that appears in the influence function of $\hat{\tau}_1$ and is computed based on observations from fold $\mathcal{I}_2$, also used in $\hat{\tau}_2$. We will now study the (asymptotic) dependence of $\hat{S}_{g,(2)}$ on $\tilde{\psi}^g_n$, the n^{th} individual influence function entering in $\hat{\pi}^g_{(2)}$. For groups g such that $\pi^g > 0$ (strong first-stage), we have that $\mathrm{P}[\hat{S}_{g,(2)} = 1] \xrightarrow[n\to\infty]{} 1$ and $\hat{S}_{g,(2)}$ becomes essentially deterministic, and therefore asymptotically there are no 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:

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

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

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

Notice that:

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

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

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

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

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

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

Hence we get:

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

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

PROOF OF LEMMA 2: The random vector $\hat{S}$ stacks the tests statistics:

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

where n^g_T denotes the test sample size in group $G=g$. 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: $\forall g \in \mathcal{G}_0,\ \pi^g = 0$. *Q.E.D.*

PROOF OF LEMMA 7: The proof follows exactly the same steps as for lemma 2 for groups with 0 and strong first-stages, however 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 is 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:

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

Q.E.D.

3. AUXILIARY LEMMAS

LEMMA 11—Influence function of the estimator of a CEF: *The influence function of the estimator* $\frac{\sum_i Z_i Y_i}{\sum Z_i}$ *of the conditional expectation function* $E[Y|Z=1]$ *is given by:* $\psi_i = \frac{Z_i(Y_i - \mathrm{E}[Y|Z=1])}{\mathrm{E}[Z]}$.

PROOF: Cf. Kennedy (2023), for instance.

Q.E.D.

LEMMA 12—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 the representations of $\hat{A}$ and $\hat{B}$ as asymptotically linear estimators into the first formula, we obtain:

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

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

Q.E.D.

LEMMA 13—Asymptotic distribution of 2SLS/Wald estimator:

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

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

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

PROOF: The Wald estimator is merely a ratio of difference of conditional expectation function (CEF) estimators and it estimates the LATE, which is a ratio of difference of CEFs. Therefore, we can see it as the combination of several asymptotically linear estimators:

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

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

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

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

$$\text{hence} \quad LATE = \frac{A-B}{C-D} \quad \text{and} \quad \widehat{\tau}^{Wald} = \frac{\hat{A}-\hat{B}}{\hat{C}-\hat{D}}$$

Since $\hat{A}, \hat{B}, \hat{C}$ and $\hat{D}$ are the estimators of conditional expectations, their influence functions are given respectively by:

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

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

We then have:

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

where (following lemma 12) $\psi_{\hat{\tau}^{Wald},i}$ is given by:

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

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

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

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

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

Q.E.D.

LEMMA 14—Bias of the test-and-select estimator in the 3-group case: *Let's consider a case with only three groups: a group with a strong first-stage ($\pi^1 > 0$), a group with a weak first-stage ($\pi^2 = H^2/\sqrt{n}$), and a group with a zero first-stage ($\pi^3 = 0$). Under assumption 3, and we have:*

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

with $B(S) = \frac{H^2 \cdot \Pr[G=2]}{\pi} \cdot \left(LATE^1 - LATE^2\right)$ *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 3.1. We can therefore 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[1] computed on the whole estimation sample. It is therefore $\sqrt{n}-$consistent (no asymptotic bias).

In the second case, the resulting estimator corresponds to the Wald estimator computed on group 1. Therefore, 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: $\pi^2 = H^2/\sqrt{n_E}$.

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

We have:

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

where the last line uses our standard notations:

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

Hence we get:

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

Therefore, we have in this case:

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

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

LEMMA 15—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 $B_w = \frac{\text{Cov}(\mu(X), w(X) | D(1) > D(0))}{\text{E}[w(X) | D(1) > D(0)]}$.

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 find:

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

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

4. DETAIL OF MONTE-CARLO SIMULATION DGPS

To simulate a flexible DGP, we use the threshold crossing model representation (Vytlacil, 2002).[2] Let $(\delta_i, \varepsilon_i)' \sim \mathcal{N}(0, \Sigma)$, with

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

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

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

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

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

We also define a covariate X as $X_i = \delta_i + \eta_i$ where $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$. X is therefore a noisy predictor of treatment receipt. Then, a group variable G is defined as the J-quantiles of X:

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

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

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

where $\tilde{\pi}_G = \pi_G - E_G(\pi_G)$ is the centered compliance rate by group with $G(i)$ representing the group G of individual i and $\nu_i \sim \mathcal{N}(0, \sigma^2_{\pi_G})$. The reason we choose this parametrization

[2] 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.

of the treatment effect is to generate a significant dependence between compliance rates and treatment effects. Indeed, with this parametrization we have:

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

so that β controls the treatment effect heterogeneity and α the dependence between the treatment effect and the compliance rate. Compared to this choice of parametrization, the one chosen in Coussens and Spiess (2021) simulation study generates very little covariance between compliance rates and treatment effects,[3] which is precisely the condition leading to a first-order bias in their estimation strategy.

Simulation to demonstrate bias in section 2 In section 2, the DGP used appends 2 DGPs

$$\text{DGP0a} \equiv \Big(N=500, J=15, S_{AT}=S_{NT}=0.5, \rho_{\delta\varepsilon}=0.9, \sigma_\delta=5, \sigma_\varepsilon=5,$$
$$\sigma_\eta=100, \alpha=0.5, \beta=10\Big)$$

$$\text{DGP0b} \equiv \Big(N=500, J=15, S_{AT}=S_{NT}=0.05, \rho_{\delta\varepsilon}=0.3, \sigma_\delta=1, \sigma_\varepsilon=1,$$
$$\sigma_\eta=100, \alpha=0.5, \beta=10\Big)$$

DGP0a creates the half of the sample with zero compliance. The large values for σ_δ and σ_ε ensure that there is large variance in the compliance rate.

DGP0b creates the other half of the population with 90% compliance and the standard parameters we use in the main simulations.

Main simulations The parameters for the DGPs in section 4 are given by:

$$\text{DGP1} \equiv \Big(N=1000, J=10, S_{AT}=S_{NT}=\frac{0.75}{2}, \rho_{\delta\varepsilon}=0.5, \sigma_\delta=1, \sigma_\varepsilon=1,$$

[3] 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.

$$\sigma_\eta = 0.01, \alpha = 0.5, \beta \in \{1, 2, 3, 4, 6, 10, 20, 40, 80\}\Big)$$

and

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

5. DETAIL OF APPLICATION

The data cleaning mainly follows Stephens and Yang (2014). We start from a sample of 1,175,889 individuals, following the sample selection of Stephens and Yang (2014). The only difference is that we do not restrict the sample to white male individuals. Stephens and Yang (2014) justify this restriction by underlining that ethnic minorities and female individuals appear to react less to compulsory schooling laws than male individuals. Instead, we suggest making such a selection in a data-driven way, starting from the full sample.

As our main covariate (G in the theory section above), we use an interaction between demographic controls (ethnicity $\times$ sex) $\times$ US census division $\times$ survey year (1960, 1970, 1980). Since we make the assumption that legal changes happen at random, we exclude from our sample the G-cells without any variation in compulsory schooling laws. Indeed, we do not want to identify the effect of compulsory schooling laws on education by comparing cells in which there has not been any legal changes with some in which there has been some, as such cells are arguably quite different. This restriction is quite stringent, and yields a sample of 171,096 individuals.

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

more of education. In other words, our treatment variable corresponds to completing some high-school education.

REFERENCES

ABADIE, ALBERTO (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113 (2), 231 – 263. [35]

ABADIE, ALBERTO, JIAYING GU, AND SHU SHEN (2022): "Instrumental Variable Estimation with First-Stage Heterogeneity," *Journal of Econometrics (forthcoming)*. [3, 5, 9, 26]

ACEMOGLU, D. AND J. ANGRIST (2006): "How Large are Human-Capital Externalities? Evidence from Compulsory Schooling Laws," *NBER Macroeconomics Annual 2000*. [30]

ANGRIST, J., G. IMBENS, AND E. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*. [6, 7]

ANGRIST, J. AND A. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*. [3]

CHERNOZHUKOV, VICTOR, MERT DEMIRER, ESTHER DUFLO, AND IVÁN FERNÁNDEZ-VAL (2021): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," *arXiv*. [34]

CLAESKENS, G AND N HJORT (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*. [35]

COUSSENS, STEPHEN AND JANN SPIESS (2021): "Instrumental Variable Estimation with First-Stage Heterogeneity," *Working Paper*. [3, 5, 22, 23, 24, 25, 26, 29, 31, 35, 60, 61, 62]

FRÖLICH, MARKUS (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139 (1), 35 – 75, endogeneity, instruments and identification. [6, 33]

HONG, HAN AND DENIS NEKIPELOV (2010): "Semiparametric efficiency in nonlinear LATE models," *Working paper*. [6]

HUNTINGTON-KLEIN, NICK (2020): "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," *Journal of Causal Inference*. [3, 5, 23, 25, 26, 29, 31]

IMBENS, GUIDO AND DONALD RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. [6]

IMBENS, G. AND E. RUBIN (1997): "Bayesian Inference for Causal Effects in Randomized Experiments with Non-Compliance," *The Annals of Statistics*. [5]

IOANNIDIS, JOHN P. A., T. D. STANLEY, AND HRISTOS DOUCOULIAGOS (2017): "The Power of Bias in Economics," *The Economic Journal*. [25]

KENNEDY, EDWARD H. (2023): "Semiparametric doubly robust targeted double machine learning: a review," . [54]

KITAGAWA, T AND C. MURIS (2016): "Model averaging in semiparametric estimation of treatment effects," *Journal of Econometrics*. [35]

LEEB, H. AND B. PÖTSCHER (2005): "Model selection and inference-Facts and Fiction," *Econometric theory*. [6, 9, 19]

OREOPOULOS, P. (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*. [30]

RUBIN, E. (1998): "More powerful randomization-based p-values in double-blind trials with non-compliance," *Statistics in Medicine*. [6]

STAIGER, E. AND J. STOCK (1997): "Instrumental variables regressions with weak instruments," *Econometrica*. [4, 8, 12]

STEPHENS, M. AND D. YANG (2014): "Compulsory Education and the Benefits of Schooling," *American Economic Review*. [3, 30, 63]

SŁOCZYŃSKI, T. (2022): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *Working Paper*. [32]

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*. [25, 61]